Université de Montréal

La webométrie en sciences sociales et humaines

Analyse des données d'usage de la plateforme Érudit

par

Sarah Cameron-Pesant

École de bibliothéconomie et des sciences de l'information

Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences

en vue de l'obtention du grade de Maître en sciences de l'information

Novembre 2016

Résumé

Cette étude exploratoire s'intéresse à l'usage des revues en sciences sociales et humaines diffusées en libre accès complet et en libre accès différé par la plateforme Érudit. Basée sur les données de téléchargements d'Érudit, elle vise à 1) fournir un portrait détaillé de l'usage des articles, 2) décrire les habitudes de téléchargement des usagers au Canada et à l'international, et 3) analyser l'effet des politiques de libre accès des revues sur les téléchargements qu'elles reçoivent. Pour ce faire, 39 437 659 téléchargements, extraits de 999 367 190 requêtes HTTP enregistrées dans les logs du serveur d'Érudit de 2010 à 2015, ont été analysés. Les résultats montrent que la majorité des usagers provient du Québec, de la France et d'autres pays francophones, et que, la plupart du temps, ceux-ci accèdent aux articles par l'intermédiaire de Google. Les habitudes de téléchargement varient d'un pays à l'autre : alors que les usagers canadiens et français utilisent Érudit principalement en journée et en semaine, leurs homologues américains sont davantage actifs en soirée, la nuit, ainsi que les week-ends. Enfin, un avantage important lié au libre accès a été observé : les articles des revues en libre accès sont davantage téléchargés que ceux des revues en libre accès différé et, pour ces dernières, la fin de l'embargo est associée à une croissance importante des téléchargements – croissance moins marquée au Canada où bon nombre d'institutions sont abonnées aux revues de la plateforme. Ces résultats démontrent l'importance des revues nationales pour les sciences sociales et humaines, ainsi que l'effet positif du libre accès sur la diffusion des connaissances, tant au Canada qu'à l'étranger.

Mots-clés : bibliométrie, webométrie, données d'usage, téléchargements, sciences sociales et humaines, Québec, Érudit

Abstract

This study explores the usage of open access (OA) and delayed OA journals in the social sciences and humanities hosted by the journal platform Érudit. Relying on Érudit's download data, the goals of the study are: 1) to describe the usage of scholarly articles, 2) to examine download patterns of national and international users, and 3) to analyze the effect of OA policies on journal download rates. The study is based on an analysis of 39,437,659 downloads, which were extracted from 999,367,190 HTTP requests stored in Érudit's log files between 2010 and 2015. The results show that the majority of users came from Quebec, France and other French-speaking countries, and that most users access articles through Google. Download patterns varied between countries: although articles were most frequently accessed during working hours, US users were more active in the evening, at night and during weekends than Canadian and French users. The study also demonstrates a clear OA advantage, as freely available articles were downloaded more frequently than delayed OA articles affected by an embargo, and downloads per article increased substantially after embargos ended. This effect was less pronounced for Canadian users, who often have access to Érudit journals via institutional subscriptions and are thus not affected by the embargo periods. The results show the positive effect of OA on knowledge dissemination in Canada as well as internationally, and emphasize the importance of national journals in the social sciences and humanities.

Keywords: bibliometrics, webometrics, usage data, downloads, social sciences and humanities, Quebec, Érudit

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles

ALM : *Article-level metrics*

APC : *Article processing charge*

CAPTCHA : *Completely automated public Turing test to tell computers and humans apart*

CRSH : Conseil de recherches en sciences humaines du Canada

CRSNG : Conseil de recherches en sciences naturelles et en génie du Canada

CSS : *Cascading style sheet*

CSV : *Comma-separated values*

FCAR : Fonds pour la formation des chercheurs et l'aide à la recherche

HTML : *Hypertext markup language*

HTTP : *Hypertext transfer protocol*

IP : Protocole Internet

IRSC : Instituts de recherche en santé du Canada

JSON : *JavaScript object notation*

NSF : National Science Foundation

OA : *Open access*

PDF : *Portable document format*

PLOS : *Public Library of Science*

RCDR : *Réseau canadien de documentation pour la recherche*

SCI : *Science Citation Index*

SciELO : Scientific Electronic Library Online

SNG : Sciences naturelles et génie

SSH : Sciences sociales et humaines

TEEAL : The Essential Electronic Agricultural Library

UNB Libraries : University of New Brunswick Libraries

URL : *Uniform resource locator*

Liste des abréviations

C.-à-d. : C'est-à-dire

Et al. : *Et alii*

Etc. : Et cætera

I. e. : *Id est*

N : Nombre de valeurs (taille de l'échantillon)

P. : Page

P. ex. : Par exemple

S. d. : Sans date

Remerciements

Ce projet n'aurait jamais pu voir le jour sans la collaboration de nombreuses personnes. En particulier, je remercie du fond du cœur Yorrick Jansen pour son aide inestimable. Il a non seulement pris en charge l'ensemble du traitement des données – incluant la détection de robots – et la conception de la première base de données relationnelle PostgreSQL, mais il m'a aussi apporté son soutien tout au long des étapes subséquentes du projet, à savoir l'interrogation des données et l'interprétation des résultats. En un mot, je lui suis reconnaissante pour sa contribution technique et intellectuelle, sa patience, sa générosité, ses précieux conseils et son support moral.

En outre, j'aimerais remercier Vincent Larivière, qui est à l'origine du projet, d'avoir cru en moi et de m'avoir donné les outils pour relever ce défi. C'est son énergie sans borne et contagieuse qui m'a convaincue de faire le passage au programme recherche – une décision dont je me réjouis. Ses nombreuses suggestions et sa grande expertise m'ont permis d'aller plus loin et d'approfondir ma réflexion. Je le remercie très sincèrement pour son soutien financier et pour les opportunités professionnelles qu'il m'a offertes.

Je tiens également à remercier tous les membres de la Chaire de recherche du Canada sur les transformations de la communication savante qui m'ont si bien accueillie. Je remercie plus spécifiquement Stefanie Haustein pour son aide et son savoir encyclopédique, ainsi qu'Adèle Paul-Hus, Antoine Archambault, Philippe Mongeon et Elise Smith pour leur écoute et leurs conseils. J'aimerais aussi exprimer ma gratitude envers les membres de l'Observatoire des Sciences et des Technologies, particulièrement envers Mario Rouette qui a repris le projet pour construire la base de données SQL Server. Enfin, l'équipe d'Érudit m'a été d'une aide considérable. Je suis spécialement redevable à Tanja Niemann pour l'accès aux données, ainsi qu'à Davin Baragiotta et David Cormier pour leur assiduité et pour leur réponse à mes nombreuses questions.

Introduction

Contexte

Histoire de la communication scientifique

Avant l'avènement des sociétés savantes, les premiers scientifiques, qui se faisaient appeler des « philosophes »[1] au XVII^e^ siècle, communiquaient de manière informelle par le biais de « collèges invisible » (« *hidden colleges* »), c'est-à-dire en s'organisant en réseaux informels de correspondance privée (Houghton, 1975, p. 12). Ces réseaux se sont progressivement formalisés en académies et en sociétés savantes dont les membres cherchaient à la fois à diffuser et à enregistrer les résultats de leurs expérimentations. Les membres de ces sociétés ont continué de communiquer par voie de correspondance et, parfois, par la publication de livres, l'arrivée de l'imprimerie dans le monde occidental au XV^e^ siècle ayant grandement contribué à la diffusion des connaissances. Cependant, les lettres ne pouvant être envoyées qu'à une personne à la fois[2] et les livres étant coûteux et longs à produire, ces sociétés ont mis au point les revues savantes comme nouveau mode de communication formel. Les correspondances entre chercheurs ont continué d'exister malgré le développement rapide des revues, mais elles sont demeurées un moyen de transmission informel des découvertes scientifiques (Manten, 1980, p. 8).

Les deux premières revues savantes ont vu le jour en 1665. Si le *Journal des Sçavans* a commencé à être publié dès janvier 1665 et que les *Philosophical Transactions* ne sont parues que le 30 mars 1665, le *Journal des Sçavans* était davantage généraliste et n'avait pas été mis sur pied pour répondre aux besoins d'une société savante (Harmon et Gross, 2007, p. 4;

---

[1] Au XVII^e^ siècle, les « philosophes », c'est-à-dire les personnes qui s'adonnaient de manière sérieuse à la science, se considéraient comme des généralistes (Manten, 1980, p. 4). Il a fallu attendre la fin du XVIII^e^ siècle pour que la science commence à se spécialiser et que l'appellation de « scientifique » apparaisse (Harmon et Gross, 2007, p. xxi).

[2] Notons que les correspondances privées n'étaient pas adressées à une seule personne, mais simplement transmises à une personne ou à un groupe à la fois. En effet, « [C]orrespondence was often distributed through the intermediary of one of the great clearing houses for scientific correspondence, such as the salon of Father Marin Mersenne in Paris, or the office of Henry Oldenburg in London. They copied the letters for further distribution. Recipients were expected to transmit the contents by word of mouth, for instance by reading the letters aloud at local meetings of scientists. » (Manten, 1980, p. 4)

Manten, 1980, p. 7). *A contrario*, les *Philosophical Transactions* servaient à publier les découvertes originales, ce qui explique qu'elles soient d'ordinaire considérées comme la première revue scientifique, malgré leur date de publication plus tardive. Ces dernières, de leur nom complet les *Philosophical Transactions: Giving some accompt of the present undertakings, studies and labours of the ingenious in many considerable parts of the world*, ont été mises sur pied à l'initiative de Henry Oldenburg, secrétaire de la Royal Society of London. Les premiers numéros des *Philosophical Transactions*, publiés mensuellement à raison de 1200 exemplaires de 20 pages, permettaient de diffuser les informations communiquées pendant les réunions de la société savante, inaugurée en 1660, tout en étant très ouverts sur le monde (Harmon et Gross, 2007, p. 2-5). Le *Journal des Sçavans*, quant à lui, a été fondé par Denys de Sallo de la Coudraye, un aristocrate parisien. À l'origine, de Sallo souhaitait publier hebdomadairement tout sujet susceptible d'intéresser les intellectuels français. Ce n'est que plus tard, en prenant le nom de *Mémoires de l'Académie royale des Sciences*, qu'il est devenu une publication officielle de cette Académie (Harmon et Gross, 2007, p. xxi). Bien que le *Journal des Sçavans* soit devenu davantage scientifique avec l'arrivée d'un nouvel éditeur, Jean Gallois, en janvier 1666, il est tout de même demeuré une publication hybride accordant la moitié de ses pages à du contenu scientifique et l'autre moitié à du contenu d'autres disciplines, telles que l'histoire, le droit, la théologie et la philosophie (Harmon et Gross, 2007, p. 40-42).

Au XVII^e^ et au XVIII^e^ siècle, de nombreux autres pays ont fondé leurs propres sociétés savantes et revues scientifiques. Pendant cette période, l'Allemagne a acquis une position dominante, presque monopolistique au sein de la science mondiale (Houghton, 1975, p. 19). Cette concentration de la production de littérature scientifique dans le nord de l'Europe s'explique par le climat plutôt libéral, la relative stabilité politique, un intérêt très répandu pour la science, la disponibilité de matériel d'imprimerie et la bonne qualité des services postaux (Manten, 1980, p. 10). Vers la fin du XVIII^e^ siècle, en réaction à la croissance de la production scientifique, les revues savantes ont commencé à se spécialiser pour ne couvrir qu'une discipline.

Au cours du XIX^e^ siècle, cette spécialisation s'est vue accompagnée de la professionnalisation de la science, c'est-à-dire de la transformation progressive de la science

en une carrière professionnelle et en une activité organisée et bureaucratique (Harmon et Gross, 2007, p. 76). Avec le début de la professionnalisation, qui s'est installée à des rythmes différents selon les pays et les langues, le nombre de publications scientifiques a commencé à croître de manière exponentielle : « During the same period, not coincidentally, the growth curve for the scientific literature shifted from linear to exponential as an intellectual gold rush spread through-out the major cities in Europe. » (Harmon et Gross, 2007, p. 76). Au cours de cette période, le nombre de scientifiques a augmenté, mais encore, ces derniers se sont mis à publier davantage, car les publications existantes ne suffisaient plus pour répondre à la demande. Les revues savantes se sont, par conséquent, multipliées et elles se sont fragmentées en différents champs de spécialisation. Dans ce contexte, les membres de la communauté scientifique se sont dotés d'une approche systématique et organisée de la communication savante. Ceux-ci, afin de remédier aux problèmes d'identification de travaux pertinents, ont donc veillé à une plus grande standardisation bibliographique et ont commencé à publier des revues de synthèse (« *abstract journals* »). En outre, un système d'évaluation plus solide que celui assuré précédemment par les éditeurs des revues ou par de petits comités éditoriaux a été développé pour garantir un meilleur contrôle de la qualité, ce qui constitue les prémisses de l'évaluation par les pairs (Meadows, 1980, p. 47-48). La deuxième moitié du XIX^{e} siècle correspond, en somme, aux débuts de la science moderne.

Une succession de découvertes scientifiques majeures a marqué le XX^{e} siècle, au fil des avancées technologiques ouvrant de nouvelles pistes de recherche. La croissance exponentielle de la production scientifique a toutefois été grandement affectée par les deux Guerres mondiales. Pour les domaines de la médecine, des sciences naturelles et de l'ingénierie, les Guerres mondiales ont eu pour effet un ralentissement considérable de la production de littérature scientifique, expliquant l'augmentation de l'âge des citations pendant ces deux périodes (Larivière, Gingras et Archambault, 2008). Cependant, depuis les années 1960, un ralentissement progressif de la croissance de la production de littérature a été constaté : « It has long been recognized in principle that the exponential growth of science at a rate greater than that of society could not continue indefinitely. » (Merton, 1973, p. 505). La distribution exponentielle s'est alors stabilisée vers une distribution linéaire, et ce, de façon encore plus marquée dès la fin des années 1970 (Larivière, Gingras et Archambault, 2008).

Selon Meadows, depuis le XVII^e siècle, les revues savantes n'ont cessé de prendre de l'importance pour la diffusion de nouvelles connaissances en science, de manière générale, et ce, malgré la diversification des moyens de communication : « the consistent trend in the development of scientific publications over the last three centuries has been for the importance of the journal to be increasingly emphasised » (1974, p. 90). La prééminence des revues a d'ailleurs aussi été accompagnée d'une augmentation importante du nombre d'éditeurs commerciaux de revues savantes, ce qui a mené à l'explosion des coûts d'abonnements aux revues depuis la fin de la Deuxième Guerre mondiale (Meadows, 1974, p. 87-88). Quoi qu'il en soit, la place occupée par les revues varie selon les disciplines. Dans les SSH, depuis 1981, les revues scientifiques prennent de plus en plus d'importance, mais la tendance est moins nette pour certaines spécialités (humanités, histoire, littérature, etc.), où le pourcentage d'articles cités a légèrement diminué depuis le début des années 1990 (Larivière, Archambault, Gingras et Vignola-Gagné, 2006, p. 1003).

Le modèle commercial de la publication savante

Situation oligopolistique

Bien qu'une présence non négligeable d'éditeurs commerciaux ait été attestée dès le XIX^e siècle, ce n'est qu'à partir de la fin de la Seconde Guerre mondiale que les grands éditeurs commerciaux ont commencé à dominer le marché de la publication savante au détriment des sociétés savantes (Larivière, Haustein et Mongeon, 2015, p. 2). Depuis la fin des années 1990, avec l'arrivée du numérique, les acteurs de l'édition savante ont commencé à s'inquiéter d'une crise : « By focusing on scholarly information as a commodity to be sold, the scientific community has been forced to enter a crisis in communication, where library funding cannot keep pace with scientific output. » (Crowther, 1999, p. [1]) Selon Guédon, cette « crise des prix des périodiques » (« *serial pricing crisis* ») (2001, p. 14-17) serait une conséquence de la création du *Science Citation Index* (SCI) par Eugène Garfield au début des années 1970. Guédon considère que le SCI, étant basé sur les revues de référence (« *core journals* ») de chaque domaine de recherche, aurait introduit malgré lui « une compétition pour un statut élitiste » dans ce qui était initialement une « quête de l'excellence » scientifique (2001, p. 17). En classant les revues les unes par rapport aux autres afin d'identifier les revues de référence

de chaque discipline, les chercheurs se sont mis à « rechercher la visibilité, le prestige, l'autorité » de ces dernières (Guédon, 2001, p. 17). C'est ce qui explique que les revues de référence, devenues indispensables pour toute bibliothèque de recherche, soient dans la mire des grands éditeurs commerciaux depuis la création du SCI.

Outre la mise sur pied du SCI, l'arrivée du numérique a mené à l'augmentation drastique de la littérature scientifique provenant de cinq grands éditeurs commerciaux, Elsevier, Taylor & Francis, Wiley-Blackwell, Springer et Sage Publications, qui publient désormais plus de la moitié de l'ensemble des articles en sciences naturelles et médecine, de même qu'en sciences sociales et humaines (SSH)[3] (Larivière, Haustein et Mongeon, 2015). Cette situation oligopolistique (Kaufman, 1998; Larivière, Haustein et Mongeon, 2015) s'explique, d'une part, par la création de nouvelles revues par ces grands éditeurs et, d'autre part, par leur acquisition massive de revues existantes. Alors que l'acquisition de revues et de petits éditeurs par de plus grandes maisons d'édition était très rare avant 1997 dans les sciences naturelles et la médecine, de même qu'en SSH, à partir de cette date, plusieurs grandes vagues d'acquisitions ont été enregistrées : en 1997 et 1998, en 2001 et en 2004 (Larivière, Haustein et Mongeon, 2015, p. 6). Le système de la publication de revues savantes s'est complexifié, les revues changeant souvent de maisons d'éditions et les maisons d'édition en acquérant de plus petites ou fusionnant avec d'autres (Larivière, Haustein et Mongeon, 2015, p. 3) De nombreuses petites revues et presses universitaires ont souffert de la crise financière qui a touché les universités dans les années 1990 ; n'ayant plus de soutien financier de la part de leur institution, elles ont dû s'associer à de grands éditeurs commerciaux (Guédon, 2001, p. 30). En outre, certains éditeurs commerciaux s'inspirèrent des archives de prépublications électroniques comme arXiv pour aider les petites revues, particulièrement celles relevant de sociétés savantes, à faire la transition vers le numérique (Guédon, 2001, p. 47). Cette situation d'oligopole s'est développée beaucoup plus rapidement dans les

[3] Les deux expressions « sciences humaines » et « humanités » sont souvent employées sans distinction dans la littérature. Historiquement, les humanités désignaient l'étude des langues et des littératures latines et grecques (Humanité, s. d.). Les sciences humaines, quant à elles, s'intéressent à l'homme et ses comportements de manière plus globale (Science, s. d.). Dans ce mémoire, nous utilisons « sciences humaines » lorsque nous désignons les sciences de l'homme de façon générale. Malgré cela, nous avons employé « humanités » lorsque nous faisons référence à un auteur qui emploie ce terme plutôt que « sciences humaines ». De plus, la classification du NSF en français que nous avons employée considère les humanités comme une spécialité de la grande discipline des sciences sociales et humaines. L'analyse de nos résultats, au chapitre 3, respecte donc cette distinction.

sciences naturelles que dans les SSH (et, parmi ces dernières, plus vite dans les sciences sociales que dans les sciences humaines), mais la création du Book Citation Index par Thomson Reuters en 2011 laisse penser que les éditeurs commerciaux s'intéressent de plus en plus à ces domaines.

L'une des raisons qui expliquent le succès rencontré par les grandes maisons d'édition commerciales est la mise sur pied d'un nouveau modèle d'affaires grâce à l'arrivée du numérique. Ainsi que l'a exprimé Derk Haank en 2001, alors président-directeur général d'Elsevier, les éditeurs ont estimé que, puisque les coûts d'accès du numérique sont moindres que ceux de l'imprimé, le prix des abonnements électroniques devraient varier selon une estimation de l'usage qu'en feront les bibliothèques universitaires (Bergstrom, Courant, McAfee et Williams, 2014, p. 9428). Ils ont donc adopté une politique discriminante de tarification. Les grands éditeurs ont, en outre, proposé aux bibliothèques une offre basée sur des grands ensembles (ou bouquets) de périodiques électroniques. L'expression anglaise « big deal » a été utilisée pour la première fois par Kenneth Frazier, bibliothécaire à l'Université du Wisconsin, très critique à l'égard de ces bouquets proposés par les grands éditeurs :

> There's no question that the Big Deal offers desirable short-term benefits, including expanded information access for the library's licensed users. In the longer run, these contracts will weaken the power of librarians and consumers to influence scholarly communication systems in the future. Librarians will lose the opportunity to shape the content or quality of journal literature through the selection process. […] The largest publishers will not only have greater market power to dictate prices. They will also have more control over contractual terms and conditions–including the ability to "disintermediate" other players in the economic chain (2001).

Selon ce point de vue, les grands ensembles ne peuvent qu'avoir des conséquences désastreuses à long terme sur l'ensemble de la communauté académique. En effet, les prix initiaux des grands ensembles ont été fixés il y a quinze ans et ne répondent plus aujourd'hui à la capacité de payer des bibliothèques universitaires, ce qui pourrait expliquer, en partie, que le nombre de bibliothèques abonnées par bouquet à l'ensemble des collections d'un grand éditeur commercial ait décru entre 2006 et 2012 (Bergstrom, Courant, McAfee et Williams, 2014, p. 9428-9429). Le coût des grands ensembles, qui augmente de 5% à 7% par année, est désormais loin d'être aussi attractif qu'auparavant. Dans le contexte où les chercheurs font pression sur les bibliothèques pour qu'elles s'abonnent aux revues jugées indispensables, les

bibliothèques universitaires n'ont que peu de marge de manœuvre pour la négociation des contrats. Plusieurs sont contraintes d'annuler des abonnements, car les grands éditeurs commerciaux les mènent à un point de rupture (Larivière, Haustein et Mongeon, 2015, p. 11).

Les revues locales, d'une grande importance pour la recherche en SSH, sont largement sous-représentées dans les bases de données bibliographiques des grands éditeurs commerciaux. La recherche dans ce domaine en est affectée, car certains chercheurs, qui souhaitent augmenter leur capital symbolique[4], peuvent choisir d'orienter leurs travaux afin de publier dans les « grandes » revues internationales (Larivière, 2014). Par ailleurs, en ce qui concerne les pays en développement, d'importants problèmes sont liés au modèle commercial extrêmement lucratif des grands éditeurs. Les universités moins nanties et n'ayant pas les moyens de s'abonner aux ressources des grands éditeurs – parmi lesquelles figurent certaines des revues les plus importantes – sont incapables de fournir une éducation supérieure de qualité ni d'exceller en recherche. Quelques pays entreprennent toutefois des démarches pour améliorer l'accès à la recherche scientifique internationale (Habib, 2010, p. 311).

La plateforme Érudit

Devant la médiatisation importante des pratiques oligopolistiques des grands éditeurs commerciaux de revues savantes, il faut toutefois se garder de condamner l'ensemble des éditeurs, une grande part de la littérature savante étant encore publiée par des éditeurs sans but lucratif. Puisque le Canada, le Québec et la France publient beaucoup en sciences sociales[5] (Boismenu et Beaudry, 2002, p. 41), et que les particularités des SSH (expliquées plus loin) amènent les chercheurs de ces disciplines à publier en grande partie dans des revues nationales, on comprend aisément l'importance de donner une visibilité aux revues nationales de langue française en SSH dans les bases de données bibliographiques, qui présentent un fort

[4] Selon Bourdieu, différentes formes de capital (économique, social, culturel, symbolique) régissent les rapports sociaux à travers les différents « champs » de l'espace social (champ littéraire, scientifique, politique, universitaire, juridique, des entreprises, religieux, journalistique [Chevallier et Chauviré, 2010, p. 21]). Le capital symbolique est le « fruit de la reconnaissance par des tiers de la légitimité de la position de celui qui en est le possesseur, et donc de sa domination » (Chevallier et Chauviré, 2010, p. 21). Autrement dit, dans un champ donné, les acteurs sont en concurrence pour maximiser leur capital symbolique, dont la valeur est propre à ce champ. Nous parlons davantage du champ scientifique dans la note 16, au chapitre 1.

[5] Plus précisément, au Canada, 75% des revues nationales sont en SSH ; au Québec, 90% des revues sont en SSH ou en arts et lettres ; en France, 90% des revues sont en SSH. En outre, 90% des éditeurs au Canada sont sans but lucratif, contre 66% pour la France (Boismenu et Beaudry, 2002, p. 41-42).

biais en faveur des revues internationales de langue anglaise. C'est dans cet esprit qu'Érudit a été mise sur pied[6].

En diffusant des revues locales et nationales dans les disciplines des SSH et des arts et lettres, la plateforme Érudit a été créée afin d'offrir aux revues savantes québécoises une plateforme de diffusion commune leur permettant de passer à l'ère numérique. Il s'agit du plus grand diffuseur de publications en SSH de langue française en Amérique du Nord (Érudit, s.d.-a). À l'Université de Montréal, « le nombre moyen de téléchargements par revue d'Érudit est plus de cinq fois plus élevé que celui d'Elsevier, douze fois celui de Wiley et 32 fois celui de Springer », ce qui démontre que « les revues nationales sont tout aussi utilisées par la communauté de cette institution que le sont les "grandes" revues internationales, et le sont bien plus que les revues publiées par la majorité des grands éditeurs » (Larivière, 2014).

Le numérique, bien qu'il ait ouvert la voie aux pratiques oligopolistiques agressives des grands éditeurs commerciaux de revues savantes, a initié d'importantes transformations dans la communication savante. En particulier, le mouvement du libre accès, en faveur de l'accès en ligne et gratuit aux publications et aux prépublications scientifiques, a ouvert un dialogue sur les enjeux liés à la diffusion des nouvelles connaissances. Une série de questions ont été soulevées par les membres de la communauté universitaire, notamment au sujet des possibilités qu'offre le format numérique pour améliorer la diffusion des nouvelles connaissances, des types de lecteurs qui devraient avoir accès aux nouvelles connaissances (p. ex. les chercheurs, le grand public, etc.), ou encore des instances qui devraient prendre en charge les frais associés à la publication électronique (p. ex. les éditeurs, les institutions universitaires, les auteurs, les lecteurs, etc.). Plusieurs modèles de libre accès ont été et continuent d'être explorés pour la publication savante. Certaines revues, éditeurs et plateformes, telles qu'Érudit, requièrent un abonnement pour l'accès aux articles courants, mais offrent les articles moins récents en libre accès. Il s'agit d'un modèle d'accès différé qui implique une période d'embargo (également appelée barrière mobile). Pour une durée d'un an, par exemple, l'accès aux articles publiés dans l'année courante est restreint, tandis que les articles publiés plus d'un an auparavant sont en libre accès. Dans le cas d'Érudit, la plupart des

[6] Le terme « Érudit » a été accordé au féminin dans l'ensemble de ce mémoire pour désigner la plateforme Érudit, y compris lorsqu'il est employé seul, sauf lorsqu'il est question du « Consortium Érudit ».

119 revues savantes[7] de la plateforme étaient soumises à une période d'embargo de 24 mois au moment de la collecte des données, mais celle-ci a été réduite à 12 mois en 2016. Il est à noter que 20 revues savantes sont en libre accès complet. Érudit diffuse également les archives de quelques revues qui ont cessé de paraître, ainsi que certaines revues culturelles québécoises[8] qui, elles, sont soumises à un embargo de trois ans[9].

Plusieurs études tendent à montrer que le libre accès augmente la diffusion et l'impact des publications scientifiques. Néanmoins, encore peu de travaux se sont intéressés spécifiquement aux SSH, où le libre accès prend davantage de temps à se répandre que dans les sciences naturelles (Eve, 2014). Les pratiques de recherche et de diffusion des connaissances variant grandement entre les domaines, il est essentiel d'étudier l'impact réel du libre accès en SSH dans le contexte où les grands éditeurs tendent à dominer le marché de la publication savante. Comme l'a démontré l'obtention d'un financement de la Fondation canadienne pour l'innovation en janvier 2015 dans le cadre du Fonds des initiatives scientifiques majeures[10], Érudit est d'une grande importance au Canada et à l'international. Cette base de données constitue donc, à nos yeux, le choix le plus indiqué pour réaliser notre étude.

Les données de téléchargements

L'une des manières d'étudier, d'un point de vue quantitatif, l'impact du libre accès sur l'usage des articles de revues savantes sur la plateforme Érudit est l'analyse des fichiers de

[7] Le rapport annuel d'Érudit fait mention de 135 revues savantes, mais les revues qui ont changé de titre au cours de leur histoire sont considérées comme des revues différentes (Érudit, 2015). Pour notre part, nous avons agrégé les différents titres des revues, ce qui donnait un total de 119 revues au moment de la collecte des données. Cela dit, de ces 119 revues, il faut savoir que les treize revues du fonds UNB (Érudit, 2016) ne sont pas enregistrées dans les logs de serveurs ni dans les données d'articles d'Érudit. Notre étude est donc basée sur les téléchargements d'articles provenant des 106 revues savantes restantes.

[8] Les revues culturelles, désignées comme telles dans les données d'articles d'Érudit, se distinguent des revues savantes en ce qu'elles s'adressent au grand public et qu'elles ne sont pas évaluées par les pairs. Elles sont, pour la plupart, recensées par la Société de développement des périodiques culturels québécois (SODEP).

[9] Notre étude se restreindra toutefois aux données de téléchargements des articles de revues savantes.

[10] Ce financement est accordé aux « installations de recherche nationales clés […] [d]e calibre mondial » (Fondation canadienne pour l'innovation, 2017). En outre, « Érudit figure parmi ces 9 installations considérées comme les plus performantes à l'échelle du pays, d'autant plus que c'est la seule soutenue pour les disciplines en sciences humaines et sociales, arts et lettres (SHSAL). Érudit fait rayonner la recherche d'ici dans le monde entier, et le financement du FCI vient reconnaitre ce caractère unique et indispensable que joue l'organisme auprès des chercheurs » (Henry, 2015). Le partenariat avec l'Agence universitaire de la francophonie montre également la vocation internationale d'Érudit.

logs de serveurs. Nous avons colligé ces fichiers logs pour la période allant du 1er avril 2010 au 31 décembre 2015. On y trouve toutes les informations relatives aux téléchargements d'articles sur la plateforme, telles que l'identifiant de l'article téléchargé, la date et l'heure du téléchargement, ainsi que l'adresse IP de l'usager. Les données de téléchargements, extraites des fichiers de logs, peuvent permettre de comparer l'usage qui est fait des articles sous embargo à celui des articles en libre accès. Or, le nombre de téléchargements comporte d'importantes limites sur le plan méthodologique dont nous reparlerons plus loin, mais qui sont trop peu explicitées dans la littérature. Répondant à un double problème de recherche, notre mémoire vise, d'une part, à étudier l'usage des articles de la plateforme Érudit, notamment l'impact du libre accès sur les téléchargements et, d'autre part, à réfléchir aux enjeux méthodologiques liés à l'analyse de données de téléchargements.

Questions de recherche

La première partie de notre mémoire est consacrée à une analyse exploratoire des données de téléchargements d'Érudit. Ce n'est qu'après avoir exploré l'usage de la plateforme et avoir pris conscience des particularités de l'indicateur du nombre de téléchargements que nous nous intéresserons à la question de l'impact du libre accès. L'analyse exploratoire nous permettra non seulement de développer un esprit critique à l'égard de ce nouvel indicateur, mais encore de mieux comprendre l'usage des articles de revues savantes sur la plateforme Érudit. Cette analyse sera elle-même divisée en deux sous-parties. Nous dresserons d'abord un portrait général de l'usage d'Érudit en tâchant de répondre aux questions suivantes :

1.1.1. Quelles sont les disciplines et les revues les plus téléchargées? Est-ce qu'elles changent avec le temps?

1.1.2. Quel est l'âge des articles téléchargés? Est-ce que l'âge moyen varie dans le temps? L'âge des articles téléchargés varie-t-il en fonction des disciplines?

1.1.3. Quels sont les pays et les continents qui téléchargent le plus? Qu'en est-il au Canada, d'une province à l'autre?

1.1.4. De quels sites Web proviennent les usagers lorsqu'ils téléchargent un article sur Érudit? Est-ce que les référents les plus courants changent dans le temps?

1.1.5. Quels sont les appareils utilisés par les usagers pour télécharger des articles (ordinateur, téléphone intelligent, tablette, etc.)? Varient-ils dans le temps?

Dans la seconde sous-partie de notre analyse exploratoire, nous accorderons une attention particulière aux habitudes de téléchargement des usagers. Nous chercherons à répondre aux questions suivantes :

1.2.1. Quelles sont les plages horaires, les jours de la semaine, les mois et les saisons où le nombre de téléchargements est le plus élevé?

1.2.2. Les habitudes de téléchargement des usagers varient-elles en fonction des pays et des disciplines?

1.2.3. Les tendances varient-elles d'une année à l'autre?

À la suite de notre analyse exploratoire, nous chercherons à mesurer l'impact du libre accès des articles sur leur usage auprès de la communauté universitaire. Deux perspectives seront adoptées : une perspective diachronique, où l'usage des articles sera observé avant et après la fin de la période d'embargo, et une perspective synchronique, où le nombre de téléchargements d'articles en libre accès sera examiné à un moment précis dans le temps. Cela nous permettra de répondre aux questions suivantes :

2.1. Le passage au libre accès semble-t-il avoir un effet sur le nombre de téléchargements? Si un effet positif est observé, quels pays, quels continents et quelles disciplines bénéficient le plus du passage au libre accès?

2.2. Si l'on tient compte du nombre d'articles en accès restreint et en libre accès disponibles dans la collection d'Érudit à un moment donné, quelle proportion des téléchargements est faite à des articles en libre accès?Notre mémoire est organisé en trois chapitres. Le premier chapitre correspond à une revue de littérature qui s'intéresse à la communication savante en SSH, à l'obsolescence de la littérature savante, à la webométrie, aux données de téléchargements et au libre accès. Le second chapitre décrit notre méthodologie, notamment les choix effectués en ce qui concerne le traitement des données, et le troisième chapitre présente nos résultats de recherche. Nous concluons notre étude par une discussion abordant les limites de l'indicateur webométrique utilisé, en espérant qu'elle saura mettre en lumière les enjeux liés à l'analyse des données de logs et contribuer à une meilleure compréhension de l'usage des articles de revues savantes canadiennes de langue française en SSH.

Chapitre 1 : Revue de littérature

La communication savante en sciences sociales et humaines

Dans cette partie de notre revue de littérature, nous aborderons brièvement les particularités des sciences sociales et humaines (SSH), puis la situation des revues dans ces disciplines au Canada.

Particularités des sciences sociales et humaines

Derek J. De Solla Price a été l'un des premiers à s'intéresser à la différence entre les pratiques de communication savante dans les sciences « dures » (« *hard science* ») et dans les sciences « molles » (« *soft science* »)[11] (1970, p. 22) Les SSH se distinguent grandement des sciences naturelles et génie (SNG), non seulement en ce qui concerne leurs objets de recherche, mais aussi en termes de pratiques de communication. En premier lieu, elles sont beaucoup plus fragmentées que les SNG. Cette fragmentation s'explique par la présence de nombreux paradigmes[12] qui s'opposent les uns aux autres dans les SSH, ce qui n'est pas le cas dans les SNG où, en général, un seul paradigme domine (Archambault, Vignola-Gagné, Côté, Larivière et Gingras, 2006, p. 332-333). L'absence de consensus entre les auteurs, que l'on constate lorsque certains auteurs évitent de citer des travaux pertinents qui ne correspondent pas à leur école de pensée (Delamont, 1989; Hicks, 1999, p. 195), a pour effet de multiplier le nombre de revues de référence utilisées par les auteurs, les auteurs adhérant à tel paradigme n'utilisant pas les mêmes revues de références que ceux adhérant à tel autre paradigme (Hicks, 1999, p. 196).

---

[11] Price a repris cette distinction entre les sciences « dures » et « molles » de Storer (1967). Elle est traditionnellement utilisée pour différencier les sciences naturelles, considérées comme plus rigoureuses, et les SSH, considérées comme davantage imprécises. Selon Storer, la rigueur scientifique, caractéristique des disciplines dites « dures », serait directement liée à l'utilisation des mathématiques dans ces disciplines.

[12] Selon Kuhn, les paradigmes sont des « découvertes scientifiques universellement reconnues qui, pour un temps, fournissent à une communauté de chercheurs des problèmes types et des solutions » (1983, p. 11). En outre, la reconnaissance d'un paradigme par les chercheurs d'un domaine donné est le signe que ce domaine scientifique entre dans la « science normale » (Kuhn, 1983, p. 30-31).

En second lieu, contrairement aux SNG où les plus importants résultats de recherche sont publiés dans des articles scientifiques, les SSH privilégient plutôt les monographies, ce qui explique le nombre important de citations faites à ce type de document. Les monographies représenteraient de 40% à 60% de la littérature en sciences sociales et elles constitueraient environ 40% des citations, toutes formes de communication savante formelle confondues (Hicks, 1999, p. 201). Aujourd'hui, la prépondérance de la monographie sur l'article savant en SSH est encore attestée dans la littérature (p. ex. Chi, Jeuris, Thijs et Glänzel, 2015; Kousha et Thelwall, 2015; Mosbah-Natanson et Gingras, 2014; Sivertsen et Larsen, 2012). À titre de comparaison, pour la Norvège, la proportion obtenue pour la période 2005-2009 est similaire à celle donnée par Hicks en 1999 : 53% des publications étaient des livres ou des chapitres de livres dans les sciences humaines et 44% des publications étaient des livres ou des chapitres de livres dans les sciences sociales (Sivertsen et Larsen, 2012, p. 570). Sachant cela, l'une des limites principales de la bibliométrie en SSH est que les monographies sont très peu indexées dans les bases de données bibliographiques.

En dernier lieu, l'objet d'étude des SSH est le contexte social dans lequel elles s'inscrivent : « En sciences humaines, l'objet de nos recherches, c'est le milieu qui nous entoure. Alors, naturellement, nous sommes portés à publier dans les revues du milieu, dans la langue que parle ce milieu […] Car c'est là que se trouve l'auditoire. » (Gingras, 1984, p. 291) Les SSH étant, par définition, davantage nationales et ancrées dans une culture spécifique, elles sont souvent diffusées dans la langue locale[13]. C'est ce qui explique l'importance des revues locales et nationales pour la recherche dans la plupart de ces disciplines (Larivière, 2014), ainsi que nous l'avons déjà mentionné. Un phénomène de globalisation et d'homogénéisation de la recherche affecte toutefois les SSH depuis une quinzaine d'années, à mesure que la pression de publier dans des revues internationales à haut facteur d'impact exercée sur les chercheurs augmente[14]. Gingras et Mosbah-Natanson

[13] Il importe de préciser que les SSH apportent aussi des contributions sur le plan international, non pas uniquement sur le plan national. Le choix de la langue de publication dépend de la « dynamique de marché » de la discipline et de la « quête de notoriété » du chercheur (Gingras, 1984, p. 291-292).

[14] Cette pression de publier dans des revues prestigieuses peut trouver des explications du côté de la sociologie des sciences. Bourdieu, notamment, décrit le champ scientifique comme « le lieu […] d'une lutte de concurrence qui a pour enjeu spécifique le monopole de l'autorité scientifique inséparablement définie comme capacité technique et comme pouvoir social […] » (1976, p. 89), ce qui va à l'encontre de l'image idéalisée de la

considèrent que ce phénomène a eu pour effet de favoriser la recherche en langue anglaise en provenance d'Europe et d'Amérique du Nord, qui était déjà dominante (2010, p. 153).

Les revues savantes en sciences sociales et humaines au Canada

En 2002, une étude basée sur huit disciplines en sciences sociales et en sciences pures a montré que la moitié des revues dominantes de ces disciplines étaient publiées par des éditeurs à but non lucratif, c'est-à-dire par des sociétés savantes et des presses universitaires (Boismenu et Beaudry, 2002, p. 12-13). C'était le cas notamment au Canada et en France, où l'édition savante se caractérisait non seulement par une présence particulièrement forte des éditeurs à but non lucratif, mais encore par une grande proportion de la recherche effectuée en SSH par rapport aux autres domaines scientifiques, par une grande autonomie dans la gestion des revues, mais aussi par une vulnérabilité financière. Bien qu'elles n'occupaient généralement pas une place dominante dans leur discipline ni dans les oligopoles internationaux, les revues nationales – et, plus particulièrement, celles en SSH et celles « s'inscrivant dans un *sous-ensemble linguistique non dominant* au plan mondial » – étaient malgré tout visibles à l'international et elles « jou[ai]ent un rôle essentiel dans la communication scientifique des diverses sociétés », ainsi que dans la publication des résultats de recherche d'intérêt plus local (Boismenu et Beaudry, 2002, p. 17-18).

En ce qui concerne les revues internationales (et dominantes) en sciences sociales, cette même étude a révélé que les éditeurs commerciaux et les éditeurs sans but lucratif se partageaient en parts égales la publication des revues (avec, pour les éditeurs sans but lucratif, 30% publiées par des sociétés savantes et 20% par des presses universitaires) (Boismenu et Beaudry, 2002, p. 32). Cependant, le prix moyen des revues d'éditeurs commerciaux était trois fois plus élevé que celui des éditeurs sans but lucratif (avec une médiane également trois

---

communauté scientifique véhiculée par les normes mertonniennes de l'universalisme, du communisme (parfois appelé communalisme), du désintéressement et du scepticisme organisé (Merton, 1973, p. 270-278). La quête de reconnaissance, dans le champ scientifique, passe notamment par la « réputation » et le « prestige » et ne peut être conférée que par des pairs – qui sont nécessairement des concurrents (Bourdieu, 1976, p. 91). Sur la notion de champ scientifique québécois, voir p. ex. Fournier, Germain, Lamarche et Maheu (1975), Fournier et Maheu (1975), ainsi que Gingras (1984, p. 289).

fois plus élevée). À qualité égale, le coût des revues de grands éditeurs commerciaux était d'autant plus grand, puisque ces derniers « ne présent[ai]ent pas un avantage particulier pour ce qui [était] de l'impact relatif de leurs revues dans la communication scientifique » (Boismenu et Beaudry, 2002, p. 37). Pour ce qui est des revues nationales, au Canada, 90% d'entre elles étaient publiées par des éditeurs sans but lucratif (Boismenu et Beaudry, 2002, p. 41).

Aujourd'hui, l'évolution de la situation des revues savantes en SSH au Canada et, plus spécifiquement, au Québec, n'a pas encore été décrite en détail. Un livre blanc préparé pour l'Association des bibliothèques de recherche du Canada indique toutefois, en 2016, qu'« [e]n raison de la diminution des sources de revenus, les presses universitaires canadiennes et l'érudition canadienne qu'elles soutiennent sont dans une situation précaire » (Whitehead et Owen, 2016, p. 14). De plus, une enquête menée par Éric Duchemin sur l'édition savante francophone en SSH au Canada, dont les résultats de la première phase ont été présentés au 84^{e} Congrès de l'Association francophone pour le savoir en mai 2016, est en cours de réalisation. Elle vise à « faire état des pratiques éditoriales et des modèles économiques en place dans l'édition savante francophone, québécoise et canadienne, numérique et papier, dans les domaines des sciences humaines et sociales » (Lebel, 2016). Enfin, un projet intitulé « L'édition savante numérique : meilleures pratiques, approches innovantes pour les revues savantes de l'Université du Québec (UQ) » est mené depuis deux ans par des chercheurs de quatre institutions de l'UQ (Caza, 2014).

Obsolescence de la littérature scientifique

Considérations générales

On pourrait définir l'obsolescence comme étant le processus de déclin de l'usage des publications scientifiques avec le temps, à mesure que le temps passe après leur publication. Le dictionnaire de Diodato en donne la définition suivante : « The decrease in use of a document or group of documents as the documents become older. Also called ageing, aging, decay. » (1994, p. 119) Dans l'ensemble, les études de l'obsolescence des articles scientifiques d'un domaine donné ont montré que la courbe de l'usage – l'usage pouvant être

mesuré au moyen de différents indicateurs bibliométriques dont nous parlerons ci-après – connaît d'abord une croissance plus ou moins importante, suivie d'un déclin plus ou moins rapide, dépendamment du domaine observé (Nicholas et al., 2005, p. 1441).

Le phénomène de l'obsolescence de la littérature scientifique est intéressant à étudier à un haut niveau d'agrégation, au niveau d'un domaine de recherche par exemple. Néanmoins, il ne reflète pas le classement des revues les unes par rapport aux autres (Glänzel et Schoepflin, 1995, p. 44). En outre, il est très peu pertinent à l'échelle individuelle étant donné le caractère non paramétrique de la distribution des citations reçues par les publications savantes. Price a d'ailleurs démontré qu'une grande proportion d'articles scientifiques n'est jamais ou que très peu citée : 35% des articles ne sont pas cités, 49% ne sont cités qu'une seule fois, 9% sont cités deux dois, 3% trois fois, 2% quatre fois, 2% cinq fois ou plus (1979, p. 158). À la suite de Price, d'autres auteurs ont confirmé que, bien que la distribution des citations varie selon les domaines, une importante proportion d'articles jamais ou peu cités côtoie une faible proportion de publications hautement citées (Seglen, 1992, p. 628) et, plus précisément, que 30% des articles ne sont jamais cités et que 20% des articles obtiennent 80% des citations (Tijssen, Visser et van Leeuwen, 2002, p. 386).

Différents modèles mathématiques ont été proposés pour décrire ce phénomène de déclin de l'usage des publications savantes depuis Gosnell, qui aurait été le premier, en 1944, à suggérer que la fonction exponentielle serait celle qui décrirait mieux à la distribution de l'obsolescence (Gosnell, 1944, p. 117; Parker, 1982, p. 129). D'autres auteurs ont par la suite avancé que l'obsolescence serait mieux décrite par une double fonction que par une seule exponentielle, notamment Meadows dans les années 1970. Selon ce dernier, la courbe de déclin du nombre de citations dans le temps est constituée de deux parties, toutes deux des exponentielles, mais à des indices différents : « one corresponding to recent literature, the other to older literature » (1974, p. 128). La première partie de la courbe, qui représente un effet de citation exagérée de la littérature scientifique récente, a été appelée « impact immédiat » (« *immediacy factor* ») par Price (1979, p. 160)[15]. Dans les années 1980, en se basant sur ses prédécesseurs, Parker a lui aussi proposé un « two-factor model » constitué de

[15] Les articles publiés pendant la période d'impact immédiat correspondent aux « publications sur le front de la recherche » (« *research front papers* ») (Price, 1979).

deux fonctions exponentielles négatives, la première appelée « *ephemeral factor* » et la seconde « *residual factor* » (1982, p. 131). Son étude montre que la fonction « éphémère » varie d'un domaine à l'autre, contrairement à la fonction « résiduelle » qui est assez stable pour tous les domaines scientifiques. Nous présenterons plus loin d'autres modèles d'obsolescence, mais basés, cette fois, sur des indicateurs webométriques plutôt que sur le nombre de citations.

Plusieurs auteurs ont constaté l'influence majeure des deux Guerres mondiales sur l'usage des publications pendant les années 1914-1925 et 1939-1950, mais les fluctuations dans l'obsolescence de la littérature scientifique au XX^e^ siècle n'avaient fait l'objet que d'hypothèses divergentes jusqu'en 2008, date à laquelle a été publiée une étude de Larivière, Archambault et Gingras. Ces derniers ont observé l'évolution de l'obsolescence pour les domaines de la médecine et des SNG de 1900 à 2004. Il s'agit d'une étude diasynchronique[16] qui compare la distribution de l'âge des articles cités à différents moments dans le temps (Larivière, Archambault et Gingras, 2008, p. 289). Des fenêtres de citations de 100 ans et de 20 ans ont été utilisées (c'est-à-dire que les références citées dans les documents étudiés ne peuvent avoir été publiées plus de 100 ou 20 ans auparavant) de manière à réduire l'effet des erreurs dans les dates de publications des références citées. Les auteurs ont conclu que :

- Plus la production de littérature scientifique est grande, moins les documents cités sont âgés et inversement, comme ce fut le cas pendant les deux Guerres mondiales où la production a diminué.
- L'augmentation de la longévité de la littérature scientifique depuis les années 1960 pourrait avoir été influencée par l'arrivée de bases de données bibliographiques.
- La science cite de la littérature de plus en plus âgée, mais certaines disciplines citent de plus en plus de littérature récente grâce aux archives de prépublications électroniques qui accélèrent la diffusion des résultats de recherche.

En ce qui concerne l'idée que la littérature scientifique aurait une plus grande longévité depuis l'arrivée du numérique, Nicholas et ses collaborateurs expliquaient, en 2005, qu'avant l'arrivée des bases de données et des outils de recherche en ligne, il était plus facile pour les

[16] Différentes approches peuvent être adoptées pour l'étude de l'obsolescence : en synchronie (c'est-à-dire que la date de citation ou la date de téléchargement est fixée, mais pas la date de publication des articles) ou en diachronie (c'est-à-dire la date de publication des articles est fixée, mais pas la date de citation ou de téléchargement) (Moed, 2005, p. 1090). Larivière, Archambault et Gingras ont utilisé une troisième approche : la diasyncronie qui compare la distribution de l'âge des articles cités à différents moments dans le temps (2008).

chercheurs d'utiliser la littérature scientifique la plus récente (2005, p. 1442). Or, de nos jours, les chercheurs peuvent aisément se référer aux documents les plus pertinents qui ne sont pas nécessairement les plus récents ; les nouvelles bases de données de recherche qui font de l'indexation rétrospective accroissent donc la visibilité de la littérature ancienne.

Depuis les années 1970, certains auteurs ont observé que l'usage et l'obsolescence des publications variaient selon les disciplines. L'usage de la littérature savante des sciences naturelles déclinerait plus rapidement que celui des sciences sociales et, à son tour, l'usage de ces dernières déclinerait plus rapidement que celui des sciences humaines dont la durée de vie est la plus grande (Houghton, 1975, p. 109-110; Line, 1993, p. 667; Nicholas et al., 2005, p. 1443). De surcroît, l'obsolescence de la littérature scientifique dans les disciplines théoriques et en recherche fondamentale est plus lente que dans les sciences appliquées (Glänzel et Schoepflin, 1999, p. 43) – ce qui pourrait sans doute expliquer le cas des mathématiques pures, où l'obsolescence de la littérature scientifique est très lente –, au même titre que l'obsolescence des publications en SSH et dans les champs professionnels est plus grande qu'en médecine et que dans les sciences naturelles (Nicholas et al., 2005, p. 1451). Plus récemment, assez peu d'auteurs se sont intéressés à l'usage des publications en SSH, mais notons Nederhof (2006), qui s'est intéressé aux limites de l'analyse des citations dans les SSH, et Chi (2016), qui a étudié, quant à lui, la concentration des citations selon les disciplines (dont les SSH) et pour différents types de documents (dont les monographies).

Indicateurs bibliométriques de l'obsolescence

Le concept de demi-vie (« *half-life* ») aurait été décrit pour la première fois par Burton et Kebler en référence au « temps requis pour que la moitié de la littérature scientifique publiée devienne obsolète » (1960, p. 19)[17]. Ce concept aurait été emprunté au domaine de la physique nucléaire pour illustrer l'obsolescence des revues scientifiques (Houghton, 1975, p. 106). On peut parler de demi-vie en termes de nombre de citations (« *cited half-life* ») pour désigner le nombre d'années de publication d'une revue, en comptant à partir de l'année en cours, pendant lesquelles cette revue a reçu 50% de son nombre total de citations, ou encore

[17] « [I]t is that time required for the obsolescence of one-half the currently published literature » (Burton et Kebler, 1960, p. 19).

de demi-vie en termes de nombre de téléchargements (« *downloaded half-life* ») pour référer au nombre de mois depuis la mise en ligne d'une revue, en comptant à partir de l'année en cours, pendant lesquelles cette revue a reçu 50% du nombre total de téléchargements d'articles (*Journal Citation Reports* de Thomson Reuters cité dans Moed, 2005, p. 1090). Or, Line explique que le concept de demi-vie n'est pas suffisant pour prédire le déclin de l'usage de la littérature scientifique :

> A further problem of accurately predicting the decline of journal use with age has been demonstrated by Maurice Line who has observed that this is a function of two factors, obsolescence and *growth*. […] Line prefers the phrase 'median citation age' to what has been used above for half-life, and substituted 'corrected half-life' which he defines as '**the half-life as estimated by removing the growth element from the median citation age**' (Houghton, 2009, p. 110) [nous soulignons].

En d'autres termes, le concept de demi-vie ne peut être utile pour décrire l'obsolescence que si l'on isole l'effet de la croissance naturelle de la production de littérature scientifique.

Aujourd'hui, de nouveaux indicateurs tels que le nombre téléchargements et le nombre de vues (ou d'accès électroniques, c'est-à-dire lorsqu'un usager accède au texte intégral, au résumé ou à des métadonnées d'un article sur le Web)[18], qui sont issus des données de logs de serveurs, constituent une mesure de plus en plus utilisée de l'obsolescence. Kurtz et ses collaborateurs ont proposé un modèle à quatre composantes pour décrire l'obsolescence de la littérature scientifique dans le domaine de l'astronomie, basé sur les accès électroniques (qu'ils appellent « *reads* » dans leur article) (Kurtz et al., 2005c). Selon ce modèle, l'obsolescence du nombre de vues peut être représentée par la somme de quatre fonctions exponentielles, que les auteurs associent à des modes d'usages différents : « historique », « intéressant », « courant » et « nouveau » (pour les numéros de revue trop récents pour être analysés) (« *historical* », « *interesting* », « *current* », « *new* ») (Kurtz et al., 2005c, p. 112). D'autres auteurs ont obtenu que l'obsolescence varie en fonction des revues et des disciplines, mais que le taux de déclin ralentit toujours avec le temps, ce qui va dans le sens de Kurtz et ses collaborateurs (Glänzel et Schoepflin, 1995, p. 44; Moed et Halevi, 2016). Basés sur les téléchargements, les résultats de Moed et Halevi tendent à montrer qu'un modèle à deux

[18] Selon Bacache-Beauvallet, Benhamou et Bourreau, il s'agit d'une mesure de l'intérêt suscité par un article ou d'une revue (2015, p. 53).

fonctions n'est pas approprié. Ils ont également observé l'obsolescence par type de document et ont constaté que les communications courtes et les éditoriaux sont les documents qui déclinent le plus rapidement, mais que le taux de déclin des articles, des articles de synthèse, des communications courtes et des éditoriaux est similaire après deux ans (Moed et Halevi, 2016, p. 418).

Des indicateurs bibliométriques aux altmetrics

Définitions

Jean Tague-Sutcliffe, dans l'introduction d'un numéro de *Information Processing & Management* portant sur l'infométrie, fait la distinction entre les champs suivants :

- L'infométrie (ou informétrie) : l'étude quantitative de tout type d'information de manière générale (c.-à-d. d'information consignée ou non).
- La bibliométrie : l'étude quantitative de la production, la diffusion et l'usage d'information consignée.
- La scientométrie : l'étude quantitative de la science (1992, p. 1).

L'expression « informétrie » aurait été, selon Brookes, suggérée par Otto Nacke en 1979 (1990, p. 35). Le terme « bibliométrie », quant à lui, a été proposé par Pritchard en 1969 pour remplacer l'expression « bibliographie statistique » (« *statistical bibliography* ») (1969, p. 348) – expression initialement utilisée par E. Wyndham Hulme dès 1922 pour désigner la discipline. Pritchard considérait toutefois qu'elle prêtait à confusion et qu'elle était peu descriptive : « Therefore it is suggested that a better name for this subject (as previously defined) is BIBLIOMETRICS, i. e. the application of mathematics and statistical methods to books and other media of communication » (1969, p. 349). Enfin, le terme « scientométrie » aurait été inventé par Dobrov et Korennoi (1969). La bibliométrie est le plus souvent appliquée à la communication savante, ce qui explique qu'elle soit parfois employée comme synonyme de scientométrie. D'autres sous-champs de l'infométrie ont aussi émergé avec l'ère numérique :

- La cybermétrie : l'étude quantitative de la production, la diffusion et l'usage de tout type d'information sur le Web (c.-à-d. d'information consignée ou non sur le Web).

- La webométrie : l'étude quantitative de la production, la diffusion et l'usage des documents Web (c.-à-d. d'information consignée sur le Web) (Thelwall, Vaughan et Björneborn, 2005, p. 84).

Comme le montre la figure 1, tirée d'un article de Björneborn et Ingwersen, le terme scientométrie peut être utilisé comme synonyme de bibliométrie, cybermétrie ou webométrie lorsqu'elle désigne l'étude quantitative de la production, la diffusion ou l'usage des publications scientifiques (sur le Web ou non). Cybermétrie et de webométrie sont parfois utilisés comme des synonymes (Björneborn, 2004, p. 12).

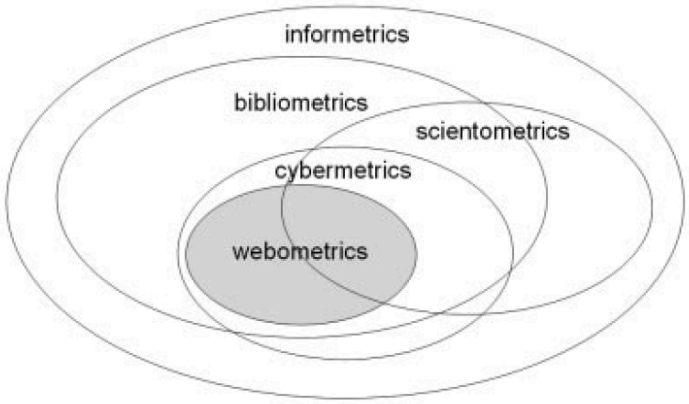

FIG. 1. Relationships between the LIS fields of infor-/biblio-/sciento-/ cyber-/webo-/metrics. Sizes of the overlapping ellipses are made for sake of clarity only.

Figure 1. Sous-disciplines de l'infométrie (Björneborn et Ingwersen, 2004, p. 1217)

À l'origine, les premiers bibliométriciens étaient issus de domaines variés tels que la bibliothéconomie, les statistiques, la sociologie ou l'histoire, et cherchaient essentiellement à étudier le système de la recherche, ainsi que ses transformations (Gingras, 2008, p. 69). Ce nouveau champ de recherche sur la science aurait vu le jour grâce à la création de la base de données de citations *Science Citation Index* (SCI) de l'Institute for Scientific Information (ISI, désormais connu sous le nom de Thomson Reuters / Clarivate Analytics), par Eugene Garfield en 1962 (Garfield, 1983; Thelwall, 2008, p. 605). L'idée lui serait venue de l'outil de recherche *Shepard's Citations,* utilisé en droit pour référencer les affaires judiciaires américaines, où pour chaque affaire sont listées les publications y ayant fait référence (Garfield, 1955, p. 108; Thackray et Brock, 2000, p. 17). En définissant les citations comme des liens formels et explicites entre des publications qui présentent certaines similitudes,

Garfield explique que les index de citations tels que le SCI sont fondés sur ces relations (1983, p. 1). À la différence de l'indexation des sujets ou des mots du titre – qui repose sur le jugement des indexeurs –, les index de citations permettent une plus grande profondeur d'indexation, moins d'ambigüité liée à la terminologie (donc une plus grande précision), des coûts réduits, une plus grande couverture temporelle et une couverture multidisciplinaire (Garfield, 1983, p. 1-5). En outre, le SCI s'appuie sur la prémisse que la majorité des articles pertinents est concentrée dans une poignée de revues de référence (« *core journals* »), ce qui est en adéquation avec la loi de Bradford[19] (Thelwall, 2008, p. 606).

Puisque le nombre de citations est un indicateur de l'impact scientifique, la bibliométrie est le plus souvent basée sur les citations. Cependant, en plus des dérives possibles (Gingras, 2014), la bibliométrie possède plusieurs limites. Selon Garfield (1983), la principale limite de la discipline serait la grande hétérogénéité des pratiques de recherche et de publication entre les disciplines. Mentionnons également les limites liées aux bases de données bibliographiques, la couverture (autant lorsque l'on vise l'exhaustivité, ce qui ne peut qu'être imparfait, que lorsque l'on souhaite constituer une sélection représentative d'un domaine), la langue, la qualité de l'indexation, etc. (Archambault, Vignola-Gagné, Côté, Larivière et Gingras, 2006; Jonkers, 2010). Ces limites sont d'autant plus importantes pour les SSH dans lesquelles les monographies ont souvent plus d'importance que les articles, et dans lesquelles une part non négligeable des publications est dans une langue autre que l'anglais, comme nous l'avons expliqué précédemment.

Webométrie

Le terme « webométrie » a été proposé par Almind et Ingwersen (1997) pour désigner l'étude quantitative de toute forme de communication basée sur un réseau – en l'occurrence, le Web. À leur suite, d'autres auteurs ont proposé des définitions légèrement différentes. Thelwall, Vaughan et Björneborn spécifient qu'il s'agit de l'étude quantitative des

[19] La loi de Bradford, nommée d'après le bibliothécaire Samuel Clement Bradford (1878-1948), est l'une des lois fondamentales en bibliométrie. Cette loi porte sur la dispersion des publications scientifiques et déclare que ces dernières sont concentrées dans un nombre restreint de revues. Plus précisément : « in a given subject field over a given period of time: (1) a few journals publish a relatively high percent of the articles in the field; (2) there are many journals that publish only a few articles each » (Diodato, 1994, p. 24).

« phénomènes reliés au Web »[20]. Dans sa thèse de doctorat, puis dans ses publications ultérieures, Björneborn définit la webométrie comme « l'étude des aspects quantitatifs de la construction et de l'usage de ressources d'information, de structures et de technologies sur le Web, employant des approches bibliométriques et infométriques »[21]. Plus encore, ce dernier explique que la webométrie regroupe quatre types d'analyse : « l'analyse de *contenu* des pages Web », « l'analyse de la *structure des hyperliens* », « l'analyse d'*usage* Web » (p. ex. l'analyse de fichiers de logs) et « l'analyse des *technologies* Web » (p. ex. l'analyse de performance des moteurs de recherche)[22] (Björneborn, 2004, p. 12).

Dès les débuts de la webométrie, Almind et Ingwersen (1997) font l'analogie entre les citations et les hyperliens, qui agissent comme les premières en faisant référence à une autre page Web[23]. Plus précisément, les hyperliens sortants (qui pointent vers d'autres pages) sont similaires aux références dans une publication, tandis que les hyperliens entrants (qui proviennent de pages extérieures et qui pointent vers une même page) s'approchent plutôt des citations reçues par une publication (Ingwersen et Björneborn, 2004, p. 340). En comparant le Web et les bases de données de citations traditionnellement utilisées en bibliométrie, Almind et Ingwersen (1997) identifient plusieurs différences, notamment la grande hétérogénéité du contenu et de la forme des pages Web qui en complexifie l'analyse. Cronin, Snyder, Rosenbaum, Martinson et Callahan appellent d'ailleurs, dès 1998, le fait de mentionner, citer ou créer un lien vers une forme de communication savante sur le Web (p. ex. une publication, une idée, un résultat, un modèle ou une théorie) une « invocation » :

> [S]cholars may find that they (i.e., their papers/ideas/findings/models/theories) have been invoked (i.e., cited/mentioned/linked to) by others. […] Polymorphous mentioning is likely to become a defining feature of Web-based scholarly communication. The phenomenon is best described by the term "invocation," which is a much broader concept than either citation or acknowledgment and better captures the multiple modalities of signaling behavior which the Web affords (1998, p. 1320).

---

[20] « [T]he quantitative study of Web-related phenomena » (Thelwall, Vaughan et Björneborn, 2005, p. 81).
[21] « The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric approaches » (Björneborn, 2004, p. 12).
[22] « [W]eb page *content* analysis », « web *link structure* analysis », « web *usage* analysis » et « web *technology* analysis » (Björneborn, 2004, p. 12).
[23] Harnad et Carr diront affirmeront même plus tard : « Bibliographic citation is the mother of all hyperlinks » (2000).

L'expression « invocation » embrasse le grand nombre de pratiques de mention sur le Web – des mentions « polymorphes » qui peuvent prendre des formes variées – et dont le contexte doit être considéré dans toute analyse webométrique.

La thèse de Björneborn, déposée en 2004, consacre un chapitre entier à la webométrie. Son objectif de recherche est de « concevoir un cadre conceptuel, ainsi que des méthodes pour identifier et décrire les phénomènes de "small-world"[24] dans le réseau d'hyperliens des espaces Web académiques »[25]. Inspiré en partie par le cadre conceptuel issu de cette thèse, Thelwall, Vaughan et Björneborn ont consacré un substantiel article de synthèse à la webométrie dans l'*Annual Review of Information Science and Technology* de 2005. Dans un article de synthèse plus récent, Thelwall précise que la webométrie inclut divers types d'analyses, notamment l'analyse d'hyperliens, l'analyse de citations sur le Web, l'évaluation des moteurs de recherche commerciaux (en particulier de leur couverture et de leur précision) et l'analyse descriptive du Web (p. ex. : la taille moyenne des pages Web, le nombre d'utilisateurs, le nombre de serveurs, etc.) (2008, p. 611). À la différence de la bibliométrie traditionnelle, qui est le plus souvent rétrospective à cause des délais de publication, la webométrie peut étudier des phénomènes très récents et elle peut porter sur des modes de communication moins formels, autres que les articles scientifiques (données de recherche, présentations, etc.) (Thelwall, 2008, p. 616). Toutefois, le contrôle de la qualité est bien plus problématique sur le Web que dans les bases de données de citations, sans compter le manque de standardisation des données. Björneborn mentionnait, en 2004, que ce champ de recherche aurait avantage à être mieux définit sur le plan théorique et méthodologique (2004, p. 23). Or, depuis le milieu des années 2000, très peu d'analyses webométriques ont été menées, en particulier en ce qui concerne l'étude de la communication savante, ainsi que le montre l'article de synthèse de Lorentzen (2014) qui ne recense aucune étude de fond dans ce champ

---

[24] « The simplest way of formulating the small-world problem is: Starting with any two people in the world, what is the probability that they will know each other ? […] While many studies in social science show how the individual is alienated and cut off from the rest of society, this study demonstrates that, in some sense, we are all bound together in a tightly knit social fabric » (Milgram, 1967, p. 62, 67). Dans son article, Milgram montre que le nombre médian d'intermédiaires nécessaires pour rejoindre une personne donnée dans le monde est de cinq ; le mode étant de six intermédiaires. C'est ce qui popularisera la notion de « *six degrees of separation* » (Guare, 1990).

[25] « The overall objective of the dissertation is to develop a conceptual framework and empirical methods concerning the identification and characterization of whether and how small-world phenomena emerge in link structures across an *academic* web space. » (Björneborn, 2004, p. vii)

de recherche depuis 2009 (Thelwall, 2009). Cela s'explique certainement par la difficulté d'obtenir l'accès aux données d'usage, rarement divulguées par les éditeurs (Haustein, 2012) – nous reviendrons sur cette question plus loin.

Altmetrics

Différents indicateurs webométriques cherchant à mesurer l'impact de la communication savante sur le Web sont aujourd'hui regroupés sous l'appellation « altmetrics » – une contraction de « alternative metrics » – que nous présenterons brièvement avant d'aborder les données de téléchargements en particulier (Priem, 2014, p. 263). Ce nouveau sous-champ de la scientométrie et de la webométrie désigne plus particulièrement « l'étude et l'usage de mesures de l'impact académique basés sur l'activité dans des outils et des environnements en ligne »[26] ; il est donc intrinsèquement lié au Web social (c.-à-d. le Web 2.0). S'il réfère souvent aux indicateurs basés sur les médias sociaux (p. ex. : Facebook), il peut aussi référer, au sens large, à d'autres mesures de l'usage de la communication scientifique sur le Web telles que le nombre de vues d'une page, le nombre de téléchargements, le nombre de mentions sur des blogs ou des micro-blogs (p. ex. : Twitter), le nombre de mentions sur Wikipédia, le nombre de sauvegardes dans un logiciel de gestion de références bibliographiques (p. ex. : Mendeley, CiteULike, Zotero), le nombre de mentions dans l'actualité et ainsi de suite.

Ce terme a été utilisé pour la première fois dans un tweet de Priem de septembre 2010 : « I like the term #articlelevelmetrics[27], but it fails to imply *diversity* of measures. Lately, I'm liking #altmetrics. » (jasonpriem, 2010) Le mois suivant[28] est paru un manifeste en faveur de l'utilisation de ces métriques prétendument « alternatives » (Priem, Taraborelli, Groth et Neylon, 2011), dont les auteurs expliquent que, dans le contexte de l'augmentation incessante du nombre de publications scientifiques, les chercheurs ont besoin de « filtres »

[26] « [T]he study and use of scholarly impact measures based on activity in online tools and environments » (Priem, 2014, p. 266).
[27] Il faut cependant préciser que les « article level metrics » et ce que l'on appelle aujourd'hui les « altmetrics » ne réfèrent pas tout à fait à la même chose, les « altmetrics » englobant les « article level metrics », mais aussi une panoplie d'« objets » de recherche (« *research objects* » (Haustein, Bowman et Costas, 2016, p. 376) qui peuvent être des agents ou des documents, ainsi que nous le verrons plus loin.
[28] Nous renvoyons à la seconde version du manifeste parue en septembre 2011, dans laquelle le tiret a été retiré de l'orthographe initiale *alt-metrics*.

pour les aider à sélectionner les travaux les plus pertinents. Selon eux, les « filtres » existants, à savoir l'évaluation par les pairs, le nombre de citations et le facteur d'impact, par exemple, sont imparfaits et ne montrent qu'une infime partie de l'impact réel de la recherche : « Unlike citation metrics, altmetrics will track impact outside the academy, impact of influential but uncited work, and impact from sources that aren't peer-reviewed » (Priem, Taraborelli, Groth et Neylon, 2011).

Le terme altmetrics est vivement critiqué, car il porte à croire que ces nouveaux indicateurs Web sont une alternative aux indicateurs traditionnels basés sur les citations, alors qu'ils en constituent plutôt un complément (Haustein, Bowman et Costas, 2016, p. 375). C'est pourquoi certains auteurs préconisent l'utilisation d'autres termes, tels que les « *social media metrics* » (Haustein, Larivière, Thelwall, Amyot et Peters, 2014). Haustein, Bowman et Costas (2016)[29] ont récemment proposé un cadre théorique ayant pour but de définir et de mieux comprendre la signification des indicateurs hétérogènes que l'on regroupe sous le terme altmetrics et qui n'ont pour point commun que leur opposition aux indicateurs bibliométriques traditionnels. Ils y décrivent trois catégories d'« actes » qui mènent à des « événements » en ligne et sur lesquels sont basés les nouveaux indicateurs : « accéder, évaluer et appliquer » (« *accessing, appraising, and applying* »), à chaque catégorie correspondant un degré d'engagement de la part de la personne qui pose l'acte (Haustein, Bowman et Costas, 2016, p. 375). Chaque type d'acte peut être appliqué à un objet de recherche, qui peut être un document ou un agent (c.-à-d. un acteur de la communauté scientifique tel qu'un chercheur, un groupe de recherche, un département, une université ou un organisme subventionnaire). Les téléchargements permettent d'« accéder » à un objet de recherche et, plus précisément, d'enregistrer un document afin de le conserver pour un usage futur (« *storing the research object* implies making it available for future use » (Haustein, Bowman et Costas, 2016, p. 378)). Les citations, quant à elles, sont à mi-chemin entre la mention et l'application.

Bornmann, quant à lui, considère que les almetrics montrent une forme d'« impact sociétal » de la science, au sens d'impact social, mais aussi culturel, environnemental et

[29] Les auteurs précisent que la nécessité de proposer un cadre théorique s'est également fait sentir dans le cas des citations, dans les premiers temps de leur utilisation.

économique de la science (2012, p. 673). Wang, Liu, Mao et Fang emploient également cette expression (2015, p. 556). En outre, les données de logs de serveurs, en particulier, permettraient de mieux comprendre l'impact social du Web selon Thelwall, Vaughan et Björneborn (2005, p. 112). Ceci dit, malgré l'effervescence des recherches depuis quelques années sur les altmetrics, on ne sait pas encore ce qu'ils mesurent réellement entre l'impact social, l'impact scientifique ou simplement un engouement éphémère (« *social impact, scientific impact or buzz* ») (Haustein, Larivière, Thelwall, Amyot et Peters, 2014, p. 10). Mingers et Leydesdorff en rappellent les principales limites :

> i) altmetrics can be gamed bv [*sic*] "buying" likes or tweets ; ii) there is little by way of theory about how and why altmetrics are generated (this is also true of traditional citations) ; iii) a high score may not mean that the paper is especially good, just on a controversial or fashionable topic ; and iv) because social media is relatively new it will under-represent older papers (2015, p. 35).

Quoi qu'il en soit, des études qualitatives devraient être menées pour mieux comprendre les raisons pour lesquelles les usagers ajoutent un marque-page, téléchargent ou partagent sur les réseaux sociaux, de même que pour évaluer si les usagers sont membres de la communauté académique ou non (Haustein, 2014, p. 339).

Les données de téléchargements

Considérations générales

Les téléchargements appartiennent à ce que l'on appelle les données d'usage, car ils permettent de mesurer une forme d'usage des publications scientifiques. Plusieurs termes sont employés dans la littérature pour désigner les données d'usage, qu'il s'agisse de téléchargements (« *hits* » [p. ex. Hitchcock, s.d.], « *web usage statistics* » [p. ex. Brody, Harnad et Carr, 2006]) ou de vues (« *hit counts* » [p. ex. Perneger, 2004], « *internet hits* » [p. ex. Perneger, 2004], « *reads* » [p. ex. Kurtz et al., 2005b], « *electronic accesses* » [p. ex. Kurtz et al., 2005b], « vues » [p. ex. Bacache-Beauvallet, Benhamou et Bourreau, 2015]). Le nombre de vues permet de mesurer le nombre de fois qu'un usager a cliqué sur un hyperlien pour parcourir les métadonnées d'un article. Il s'agit davantage d'une mesure de l'intérêt suscité par un article que de son usage, car l'usager ne télécharge pas nécessairement le texte

intégral de l'article (Bacache-Beauvallet, Benhamou et Bourreau, 2015, p. 53). Le nombre de téléchargements[30], quant à lui, est une mesure plus précise de l'usage d'un article scientifique dans un champ de recherche donné, mais il demeure tout de même un indicateur imparfait dont il importe de connaître les limites.

Kurtz et Bollen proposent une définition de l'usage qui souhaite tenir compte de l'ensemble des contextes possibles : « [u]sage occurs when a user issues a request for a service pertaining to a particular scholarly resource to a particular information service » (2010, p. 6). Dans cette définition, ils excluent les variables non-mesurables telles que les intentions de l'usager. Ils distinguent ensuite la notion générale d'« usage » de celle d'« événement » et de « données d'usage » :

> A usage event is the electronic record of a user-generated request for a particular resource, mediated by a particular information service, at a particular point in time. Usage log data are collections of individual usage events recorded for a given period of time (Kurtz et Bollen, 2010, p. 6).

En d'autres termes, les données d'usage (p. ex. enregistrées dans des logs de serveurs) sont un ensemble d'événements individuels (p. ex. des téléchargements) au cours desquels des usagers font une requête pour obtenir un document[31]. De leur côté, Moed et Halevi adoptent une définition plus pragmatique de l'usage : « the use made of electronic publication archives in the broadest sense, and recorded in the archive's electronic log files » (2016, p. 413). Ils distinguent l'usage de l'utilisation (« *use* »), un terme neutre qui s'applique à la fois à l'usage (à savoir l'usage d'une publication savante sous forme électronique) et aux citations (reçues par cette publication) (Moed et Halevi, 2016, p. 414).

Plusieurs auteurs considèrent les téléchargements et les citations comme deux aspects ou deux étapes différentes dans un même processus de diffusion des connaissances. Selon Moed, il s'agit de « phases distinctes dans le processus de collecte et d'analyse de l'information scientifique pertinente qui peut éventuellement mener à une nouvelle

[30] Certains auteurs utilisent les données de téléchargement du résumé, de la table des matières et/ou du texte intégral d'articles scientifiques, mais notre analyse ne s'intéressera qu'aux téléchargements du texte intégral des articles savants sur Érudit.

[31] Les données d'usage se distinguent des statistiques d'usage. Ces dernières sont des statistiques tirées des données d'usage ; elles sont donc dépourvues de toute information liée à un événement individuel, contrairement aux données d'usage (Kurtz et Bollen, 2010, p. 7-8).

publication »[32] (2005, p. 1096). Il s'inspire également de la distinction de Garvey et Griffith[33] pour préciser que les téléchargements représentent une forme d'usage informel de documents formels (c.-à-d. les publications scientifiques), tandis que les citations représentent l'usage formel de documents formels (Moed, 2005, p. 1089). Priem et Hemminger, quant à eux, considèrent que les données d'usage permettent de mesurer « l'impact sur les lecteurs », à la différence des citations qui mesurent « l'impact sur les auteurs »[34] (2010). Dans le même ordre d'idées, Haustein, considère les téléchargements comme un indicateur portant sur le lectorat (« *readership metrics* ») : « A more accurate way of analyzing scholarly journals is to consider being read and being cited as two different aspects of influence, and thus journal usage and journal citations as two separate dimensions of journal evaluation » (2014, p. 329). Kurtz et ses collaborateurs estiment que le nombre d'accès électroniques au texte intégral d'articles scientifiques constitue un nouvel indicateur aussi important que les citations pour mesurer l'impact d'un article scientifique (2005b, p. 36). Un sondage de Rowlands et Nicholas mené en 2005 à l'international a montré que les chercheurs séniors considèrent le nombre de téléchargements comme un meilleur indicateur de l'« utilité » (« *usefulness* ») de la recherche que les citations (2005, p. 27). Par ailleurs, l'un des avantages évident des téléchargements est qu'ils peuvent être observés beaucoup plus rapidement que les citations qui, elles, ne peuvent être mesurées qu'après plusieurs années (Perneger, 2004, p. 546).

Moed et Halevi, dans une récente étude, proposent une liste fort utile – que nous traduisons et paraphrasons ici – de dix points permettant de différencier les téléchargements des citations (2016, p. 6-8) :

1. Les logs de serveurs ne permettent pas de repérer toutes les sources où repêcher une publication (p. ex. une prépublication ou une autre version en libre accès, une copie partagée par un collègue, un autre point d'accès sur le Web, etc.).

---

[32] « [D]ownloads and citations relate to distinct phases in the process of collecting and processing relevant scientific information that eventually leads to the publication of a journal article » (Moed, 2005, p. 1096).

[33] Dans le domaine de la physique, Garvey et Griffith décrivent les échanges informels comme étant tout ce qui vient avant la publication dans une revue savante, par exemple : discussions entre collègues, rapports de résultats préliminaires, communications, actes de congrès, prépublications, manuscrits soumis et rejetés, etc. (1971, p. 353). Les échanges formels comprendraient, notamment : articles scientifiques, revues de synthèse, revues annuelles, bulletins de sociétés savantes, citations, etc.

[34] « The migration of academic literature to the Web allows us to examine views or downloads for most articles; instead of measuring an article's impact on authors (who may or may not cite it), usage data supports measurement of impact on readers » (Priem et Hemminger, 2010).

2. Les bases de données de citations présentent d'importants problèmes de couverture, notamment pour ce qui est des livres, des revues nationales et des publications dans des langues autres que l'anglais.
3. Un téléchargement ne mène pas toujours à une lecture complète de l'article, ce qui est moins souvent le cas pour les citations.
4. Les usagers qui lisent (les lecteurs) et ceux qui citent (les auteurs) n'appartiennent pas toujours au même groupe dans la communauté académique. Les lecteurs peuvent être des chercheurs, des professionnels, des étudiants de premier cycle ou des intéressés parmi le grand public, mais ce sont les chercheurs qui publient et citent.
5. Le nombre de téléchargements, ainsi que le ratio entre le nombre de téléchargements et le nombre de citations varient selon le type de document. Par exemple, les éditoriaux sont beaucoup téléchargés, mais peu cités comparativement aux articles.
6. L'obsolescence du point de vue des téléchargements peut être observée très peu de temps après une publication, à la différence de l'obsolescence du point de vue des citations qui ne peut être observée qu'après quelques mois ou quelques années.
7. Les téléchargements sont un indicateur de l'attention reçue par un article ou de l'intérêt d'un usager pour un article, tandis que les citations sont un indicateur de la réflexion faite par un auteur sur un article pendant ses recherches.
8. Les téléchargements et les citations peuvent s'influencer mutuellement : un téléchargement peut mener à des citations et inversement.
9. Le nombre de téléchargements peut être manipulé beaucoup plus facilement que les citations, ces dernières étant soumises à l'évaluation par les pairs.
10. Contrairement aux citations qui sont « publiques », les téléchargements peuvent impliquer des enjeux de vie privée lorsqu'ils sont agrégés au niveau des individus, des institutions ou des organismes subventionnaires.

Cette liste de caractéristiques aide à mieux comprendre comment utiliser et interpréter l'indicateur des téléchargements et celui du nombre de citations. Elle pose les bases des principaux avantages et limites des téléchargements dont nous reparlerons plus loin.

Corrélation entre le nombre de téléchargements et le nombre de citations

Depuis le début des années 2000, plusieurs auteurs ont obtenu une corrélation entre le nombre de téléchargements et le nombre de citations que recevra un article plus tard, notamment Hickman (2000)[35], Harnad et Carr (2000)[36], Brody, Carr et Harnad (2002)[37],

[35] « [T]here is a correlation between downloads and citations and […] highly cited papers get downloaded more » (Hickman, 2000).

Hitchcock et al. (2002)[38], Hitchcock, Brody, Gutteridge, Carr et Harnad (2003)[39], Perneger (2004)[40], Jahandideh, Abdolmaleki et Asadabadi (2007)[41], O'Leary (2008)[42], Lippi et Favaloro (2012)[43], ou encore Kurtz et Henneken (2016)[44]. En outre, certains ont constaté que plus la période pour laquelle le nombre de téléchargements compté est longue, plus forte sera la corrélation entre les citations et les téléchargements (Brody, Harnad et Carr, 2006; Moed, 2005), Moed mettant toutefois en garde contre l'interprétation d'une telle corrélation comme étant une relation de causalité.

L'étude récente de Moed et Halevi a pour objectif de mieux comprendre les différences dans la distribution des téléchargements et dans celle des citations grâce à l'analyse statistique d'une grande variété de questions sur le sujet (2016). Parmi les résultats les plus intéressants qu'ils ont obtenus, notons :

- Le ratio du nombre de téléchargements et du nombre de citations varie selon les domaines, mais les SSH tendent à avoir des ratios très larges.
- La distribution non paramétrique des citations tend à être plus concentrée que celle des téléchargements parmi les articles d'une même revue.
- À l'échelle des revues et des articles, la corrélation entre les téléchargements et les citations est forte dans les SNG, mais faible dans les SSH.

---

[36] « After the initial download peak, the (eventually) more highly cited papers show higher and more sustained download frequency » (Harnad et Carr, 2000).
[37] « It is possible to see a correlation between papers that are highly cited […] and those which are frequently downloaded. It is not known whether frequently downloaded papers lead to more citations, or whether the citations […] leads to more downloads. It is demonstrable that highly cited papers have a higher download longevity–they are downloaded more for longer » (Brody, Carr et Harnad, 2002, p. 75).
[38] « [T]he peak of citations occurs higher and sooner for papers deposited in each succeeding year […] [and] high impact papers are accessed more often and over a more sustained period » (Hitchcock et al., 2002).
[39] « [U]sage impact is correlated with citation impact, i.e. the more often a paper is downloaded the more likely it is to be cited. This correlation is highest for high-citation papers and authors » (Hitchcock, Brody, Gutteridge, Carr et Harnad, 2003).
[40] « The hit count was associated with the number of subsequent citations » (Perneger, 2004, p. 547).
[41] « [M]ore downloads at a limited period of time is an indicator of more citations to the article in a long-term interval » (Jahandideh, Abdolmaleki et Asadabadi, 2007, p. 459).
[42] « [T]his paper finds [a] strong positive statistically significant relationship between the number of citations and downloads of papers in *Decision Support Systems*. […] Papers among the most downloaded papers receive a higher average number of citations than "normal" papers […] » (O'Leary, 2008, p. 979).
[43] « [T]he most downloaded articles in the field of laboratory medicine are also those that are more likely to receive citations in the short term (i.e., in the following 2 years) » (Lippi et Favaloro, 2012, p. 195).
[44] « [D]ownloads do, indeed, predict citations » (Kurtz et Henneken, 2016, p. 8).

- Les articles les plus téléchargés d'une revue (plus de 2 000 téléchargements) ont tous reçus un minimum de dix citations, tandis que les articles les plus cités (plus de 20 citations) ont tous été téléchargés un minimum de 500 fois.

Les auteurs proposent l'hypothèse que les domaines où la corrélation est la plus forte semblent être ceux dans lesquels le lectorat se limite à des chercheurs actifs, tandis que les domaines où la corrélation est la plus faible semblent être ceux dans lesquels le lectorat peut inclure le grand public et des professionnels (Moed et Halevi, 2016, p. 429). En conclusion, ils comparent leurs résultats avec ceux obtenus par Kurtz et ses collaborateurs (2005c). Ces derniers avaient conclu que « les citations sont un bon prédicteur des téléchargements », contrairement aux « téléchargements [qui] ne prédisent pas les citations »[45] (Kurtz et al., 2005c, p. 117). Or, les résultats de Moed et Halevi ne permettent pas d'arriver à une telle conclusion : « [t]here is perhaps even more evidence for the reverse conclusion, namely, that downloads are a good predictor of citations and citations a poor, or in any case a less valid, predictor of downloads » (2016, p. 427).

Si plusieurs auteurs ont obtenu une certaine corrélation entre le nombre de téléchargements et le nombre de citations des articles, d'autres auteurs plus prudents constatent que les corrélations obtenues sont bien trop faibles pour affirmer que les téléchargements et les citations mesurent la même chose (Haustein, 2014, p. 331; Li, Thelwall et Giustini, 2011, p. 3; Thelwall, 2008, p. 611). L'hétérogénéité des résultats obtenus en matière de corrélation entre ces deux indicateurs laisse penser que les données d'usage montrent une forme d'impact différente de celle des citations. Cela dit, dans leur plus récent article, Kurtz et Henneken sont parvenus à la conclusion que les citations et les téléchargements sont des mesures aussi efficace l'une que l'autre pour mesurer et même prédire la performance de chercheurs individuels (2016). Plus précisément, ils ont comparé l'efficacité des mesures de citations et de téléchargements pour évaluer la performance passée, présente et potentielle d'un échantillon d'astrophysiciens à différents moments de leur carrière. Leurs résultats montrent que les téléchargements prédisent bel et bien les citations, mais uniquement dans le cas où l'on parvient à isoler les téléchargements effectués par des

[45] « The number of citations is thus a good predictor of the number of reads […].The same cannot be said of the number of reads; they clearly make a very poor predictor of the number of citations » (Kurtz et al., 2005c, p. 117).

chercheurs. Cependant, puisque ces mesures ont des propriétés différentes[46], elles peuvent pour cette raison donner des résultats différents pour un même chercheur. Les auteurs recommandent d'utiliser à la fois les citations, les téléchargements et l'évaluation par les pairs pour évaluer les chercheurs à l'échelle individuelle.

Habitudes de téléchargement des usagers

Les enjeux liés à la conciliation entre la vie professionnelle et la vie privée sont abordés depuis la révolution industrielle, mais ils sont l'objet de vives préoccupations depuis les dernières décennies à cause des exigences de plus en plus élevées sur le marché du travail (Guest, 2002, p. 256). C'est le cas notamment dans le monde académique, où la culture de « *publish or perish* » prend davantage d'ampleur à mesure que la compétition s'intensifie pour l'obtention du financement et de l'agrégation. Selon Harnad, le phénomène de « publier ou périr » comporte trois facettes : le besoin croissant des chercheurs à publier un nombre élevé d'articles, la nécessité de publier dans des revues de qualité qui ont une bonne réputation et l'importance que les articles publiés aient un impact pour l'avancement des connaissances (2011, p. 34). Dans ce contexte, de nombreux auteurs ont commencé à s'intéresser aux habitudes de travail des chercheurs, notamment sous l'angle de la conciliation entre la vie professionnelle et la vie privée en sciences de l'information (Cabanac et Hartley, 2013), des effets de calendrier sur la diffusion de la science (Magnone, 2013), de l'emploi du temps des chercheurs (Wang et al., 2012), ainsi que des influences saisonnières et des cycles de vie académiques d'une année à l'autre (Moed et Halevi, 2016).

Wang et ses collaborateurs (2012) ont analysé l'emploi du temps des chercheurs à travers le monde au moyen de téléchargements. Cette étude présuppose que les téléchargements sont un indicateur valable d'une certaine forme de dévouement au travail :

> Admittedly, working involves many kinds of behaviors, but it is certain that if someone is downloading literatures, he is definitely working at the table. Thus, to explore scientists' working habits reflected by working hours, we have been conducting this research on the downloads [*sic*] of scientific literatures (Wang et al., 2012, p. 655).

[46] « [Citation] measures are always weighted towards the first half of the individual's career [while] [d]ownloads have the opposite problem, they principally measure recent performance » (Kurtz et Henneken, 2016, p. 9).

L'échantillon comprend les téléchargements de quatre jours de la semaine et de quatre jours de week-end au mois d'avril 2012, pour un total de plus de 1 800 000 téléchargements. Les auteurs ont comparé les périodes d'activités, basées sur l'heure locale, des trois pays qui téléchargent le plus sur Springer, à savoir les États-Unis, la Chine et l'Allemagne. Les résultats montrent que les chercheurs américains sont ceux qui téléchargent le plus durant la nuit, autant en semaine que le week-end. En effet, le nombre de téléchargements augmente progressivement de 7h à 16h, puis redescend, sans que l'heure des repas ne soit visible. Pour ce qui est de l'Allemagne, le nombre de téléchargements augmente rapidement de 7h à 11h, redescend à partir de 12h en semaine, remonte à 13h, puis décline en dents de scie jusqu'au lendemain matin. Il y a très peu d'activité pendant la nuit. En Chine, le nombre de téléchargements augmente très rapidement de 7h à 11h. Il y a deux importants creux autour de 12h et de 18h à cause des repas servis à heure fixe sur les campus, de même que trois pics à 10h30, 15h30 et 21h. Pour les trois pays, les téléchargements sont assez élevés le week-end, mais la Chine l'emporte avec des chercheurs très actifs le week-end, en particulier en après-midi et en soirée (Wang et al., 2012, p. 658).

L'étude de Magnone (2013), pour sa part, est basée sur trois jeux de données d'articles savants sur la plateforme ScienceDirect d'Elsevier : les articles publiés entre 1990 et 2010, les articles publiés en 2008 et les articles publiés en 2008 soumis par des auteurs chinois. Ses résultats montrent que le nombre de soumissions est fortement influencé par les événements du calendrier, tels que les fêtes, les jours fériés ou les vacances. Selon l'auteur, la date de soumission des articles est beaucoup plus représentative du dévouement au travail des chercheurs que la date de publication, car cette dernière ne dépend pas de l'auteur, mais de la revue. Parmi les résultats obtenus, soulignons une forte diminution du nombre de soumissions à la fin décembre ; une augmentation du nombre de soumissions au printemps et à l'été, suivie d'une diminution à l'automne et à l'hiver ; la somme des soumissions en semaine est sept fois plus élevée que la somme du week-end, le maximum étant atteint le mardi ; le nombre de soumissions est plus élevé au milieu du mois qu'au début et qu'à la fin[47] ; une forte

[47] Magnone justifie ce résultat de la manière suivante : « the only possible explanation is that most of the scientific projects ended at the end of every single month » (Magnone, 2013, p. 106). Néanmoins, il ne précise pas sur quoi il se base pour affirmer cela.

diminution du nombre de soumissions le 25 décembre, le 1er janvier et le 7 février 2008 (c.-à-d. la date du nouvel an chinois).

Cabanac et Hartley (2013), quant à eux, se sont intéressés à la question de la conciliation entre la vie professionnelle et la vie privée des auteurs et des éditeurs de la revue *JASIST*. Ils s'inscrivent notamment dans la lignée de Wang et ses collaborateurs, ainsi que de Magnone, tout en insistant sur les trois importantes limites de ces études : aucune technique de détection de robots qui peuvent influencer les tendances ne semble avoir été mise en place ; les résultats obtenus sous-estiment sans doute la quantité de travail effectué par les chercheurs le week-end, car un téléchargement en semaine peut mener à une lecture le week-end ; l'échantillon de Wang et al. est jugé beaucoup trop petit et pourrait avoir été biaisé par des événements spéciaux[48]. Pour pallier ces limites, Cabanac et Hartley ont basé leur analyse sur l'historique de publication des articles de la revue *JASIST* (c.-à-d. la date où un article a été soumis pour la première fois, la date où il a été révisé, de même que la date où il a été accepté) plutôt que sur les téléchargements, car ils estiment que l'historique de publication représente mieux le dévouement au travail des chercheurs (2013, p. 2). Les auteurs de *JASIST* n'ont aucune date limite pour soumettre un article et ils ont ensuite une année entière pour le réviser, c'est donc leur choix de travailler les week-ends. Ils soumettent le plus souvent leurs articles en semaine, mais les week-ends obtiennent tout de même 11% des soumissions totales (autant pour les nouvelles soumissions que pour les soumissions finales). En ce qui concerne les éditeurs, les articles acceptés le week-end constituent 7% des acceptations totales. Jusqu'à 2004, les auteurs de *JASIST* étaient peu actifs les samedis et dimanches, mais le nombre de soumissions a augmenté de 1% par année, jusqu'à atteindre 20% de l'ensemble des soumissions en 2012. Cabanac et Hartley spécifient que l'une des limites de leur étude est que, pour des raisons techniques, ils n'ont pas pu tenir compte des fuseaux horaires pour déterminer la date locale des soumissions, même si seulement 50% des articles de *JASIST* proviennent des États-Unis.

[48] Nous ajouterions aux reproches effectués à Wang et ses collaborateurs le fait d'avoir exclu une journée entière de leur échantillon (ils sont passés de neuf à huit jours) à cause de valeurs jugées anormales, plutôt que de chercher à identifier le problème. Ils ne définissent pas non plus les limites de ce qu'ils considèrent anormal.

L'étude la plus récente qui aborde la question des habitudes de téléchargements des chercheurs est celle de Moed et Halevi (2016)[49]. Les auteurs ont analysé l'influence des saisons et les cycles de vie académiques par mois de l'année pour la Chine et le Royaume-Uni et ont observé des cycles se répétant annuellement de 2009 à 2013.

Limites

L'accès aux fichiers de logs constitue souvent un premier obstacle à surmonter pour mener une étude sur les téléchargements ; ceux-ci étant rarement disponibles publiquement[50]. Des problèmes de confidentialité peuvent d'ailleurs se poser quant à l'accès à certaines informations qui sont parfois enregistrées dans les logs, ou encore dans le cas des « *deep log analysis* », une méthodologie qui consiste à croiser les données de logs avec des données sociodémographiques des usagers (Huntington, Nicholas, Jamali et Tenopir, 2006; Nicholas et al., 2005). L'analyse des logs peut également poser de grandes difficultés pour ce qui est d'identifier les utilisateurs et les sessions de recherche, de distinguer les requêtes effectuées par des humains de celles effectuées par des robots[51], de localiser géographiquement les requêtes et de faire le traitement d'un énorme volume de données (Priem et Hemminger, 2010; Thelwall, Vaughan et Björneborn, 2005).

Bien que la majorité des auteurs s'entende pour dire que les téléchargements sont un indicateur de l'impact ou de l'influence des publications scientifiques sur les lecteurs, le terme « *reads* », employé par Kurtz et ses collèges (2005b; 2005c) peut porter à confusion, car un téléchargement ne mène pas toujours à une lecture[52] et un document téléchargé n'est pas nécessairement lu par la personne l'ayant téléchargé (Davis, Lewenstein, Simon, Booth et Connolly, 2008, p. 5). Plusieurs raisons peuvent justifier le téléchargement d'un article : faire une recherche exploratoire dans une base de données, être abonné à une liste de diffusion, consulter une référence, etc. (Moed, 2005, p. 1088). Brody, Harnad et Carr ajoutent qu'un

[49] Nous ne parlons que brièvement de cette étude puisque la question des habitudes de téléchargements des chercheurs demeure, somme toute, assez périphérique.
[50] Les revues de la *Public Library of Science* sont parmi les rares à diffuser des données sur les téléchargements des articles qu'ils publient : https://plos.org/article-level-metrics
[51] Nous aborderons la détection de robots en détails dans le chapitre 2 – Méthodologie, dans la section portant sur le traitement des données.
[52] Il arrive d'ailleurs parfois que ce soit le cas pour les citations lorsqu'un auteur reprend une citation de seconde main (Davis, Lewenstein, Simon, Booth et Connolly, 2008, p. 5).

usager peut télécharger plusieurs fois le même article, que de nombreux usagers peuvent télécharger un article derrière un proxy ou une adresse IP unique, qu'un usager peut distribuer plusieurs copies imprimées à des collègues et, enfin, qu'un même article peut être téléchargé depuis différentes sources (p. ex. des archives de prépublications électroniques, le site Web personnel de l'auteur, etc.) (2006, p. 1063). Certains auteurs mettent d'ailleurs en garde contre l'analyse des téléchargements effectués au cours des sept jours suivant la publication d'un article, car ils reflètent simplement l'effet des alertes par courriel (Brody, Harnad et Carr, 2006; Hickman, 2000).

Le libre accès

Considérations générales

Dans le domaine de la communication savante, le mouvement du libre accès prône la diffusion la plus large possible des publications scientifiques grâce aux possibilités offertes par le format numérique. Eve propose la définition suivante :

> The term 'open access' refers to the removal of price and permission barriers to scholarly research. Open access means peer-reviewed academic research work that is free to read online and that anybody may redistribute and reuse, with some restrictions (2014, p. 1).

Le prix constitue une barrière à l'accès fortement dissuasive pour les bibliothèques universitaires qui doivent être très sélectives dans le développement de leurs collections de périodiques, de même que pour les chercheurs qui doivent consulter une quantité parfois importante de travaux auxquels leur institution n'est pas abonnée. Le copyright peut également constituer une barrière pour la traduction d'une publication dans une autre langue, la distribution de copies à des collègues, la fouille de textes ou encore la migration vers un nouveau format (Suber, 2012, p. 5). Le libre accès, en assouplissant les droits de propriété intellectuelle, élimine les deux principales barrières à l'accès[53].

[53] Malgré cela, le libre accès ne rend pas universel l'accès aux publications scientifiques, car quatre types de barrières demeurent : la censure de gouvernements, d'institutions ou d'individus, la barrière de la langue, l'accès pour les personnes handicapées et l'accès à une connexion Internet (Suber, 2012, p. 26-27).

Plusieurs supports peuvent véhiculer le libre accès, mais les deux principaux sont les revues savantes et les répertoires (répertoires institutionnels ou disciplinaires, archives ouvertes, archives de prépublications électroniques ou toute autre forme de base de données d'articles scientifiques en ligne [Suber, 2012, p. 49-52]).

Ce que l'on appelle la « voie dorée » fait référence à toutes les formes de libre accès par le biais des revues savantes, peu importe le modèle d'affaires de celles-ci. Les revues peuvent être entièrement en libre accès ou elles peuvent être hybrides, auquel cas elles contiennent à la fois des articles en accès restreint (c.-à-d. sur abonnement payant) et des articles en libre accès. Elles peuvent exiger des frais de la part des auteurs, de leur organisme subventionnaire ou de leur institution afin de couvrir les coûts de publication ; ce sont les « *article processing charge* » (APC) (Eve, 2014, p. 179). Le libre accès différé, dans le cas où des articles en accès restreint deviennent en libre accès après une certaine période de temps, nommée barrière mobile ou embargo, est l'une des nombreuses formes que peut prendre la voie dorée. Parmi les autres formes de libre accès doré, mentionnons aussi la voie « *freemium* », c'est-à-dire l'« accès ouvert au texte html » conjointement avec la « vente de services associés comme la fourniture de PDF téléchargeables et réutilisables à volonté, ou fourniture de statistiques de consultation pour les bibliothèques » (Contat et Gremillet, 2015, p. 12), ainsi que la « voie platine »[54] qui consisterait :

> à voir les revues de S.H.S. [sciences humaines et sociales] financées 'en amont', pour partie par le Ministère, pour partie par les établissements, dans le cadre d'un accord national, en échange de l'engagement des éditeurs de ces publications de les diffuser gratuitement sous forme électronique, sans aucune période d'embargo (IDATE et Cairn.info, 2015b).

La « voie verte », quant à elle, correspond au libre accès par le biais de dépôts institutionnels, disciplinaires (tels *arXiv*) ou de sites Web personnels. Synonyme d'« auto-archivage », il s'agit, pour un auteur, de déposer la version finale accepté d'un document ou une prépublication sur un répertoire en ligne ou sur son propre site. Les répertoires peuvent contenir des prépublications, des publications évaluées par les pairs, de même que d'autres types de documents tels que des thèses, des mémoires, des jeux de données, des livres, des

[54] Eve considère toutefois qu'il s'agit d'un terme erroné sensé désigner une forme de libre accès doré sans APC, mais qui n'est, en fait, que l'un des modèles possibles de libre accès doré (2014, p. 181).

chapitres de livres, des documents numérisés, etc. Les répertoires institutionnels visent à rassembler la production scientifique d'une institution, tandis que les répertoires disciplinaires réunissent celle d'une discipline ou d'un champ de recherche. L'idée du libre accès par le biais d'archives ouvertes a été grandement diffusée par la « proposition subversive » de Stevan Harnad du 27 juin 1994, dans laquelle il dénonce violemment la vente des publications savantes et propose comme solution la création d'un serveur FTP public (« *a globally accessible local ftp archive* » [Okerson et O'Donnell, 1995, p. 11-12]).

Une prépublication est un document qui présente des résultats de recherche, mais qui n'a pas encore été évalué par les pairs, contrairement aux documents publiés dans des revues savantes (Boismenu et Beaudry, 2002, p. 27-28). Elle peut faire l'objet d'une publication *a posteriori* ou non. En physique des particules, notamment, les prépublications sont très utilisées afin d'accélérer la diffusion des résultats de recherche les plus récents (Thelwall, Vaughan et Björneborn, 2005, p. 102), ce qui explique la création de l'archive de prépublications électroniques *arXiv* par Paul Ginsparg dès 1991 (Ginsparg, 2011). Les prépublications ne visent pas à remplacer les revues (Couture, 2013) et ne leur nuisent pas non plus, comme l'a montré l'exemple d'*arXiv* qui n'a causé aucun tort à la revue *Physical Review* (Boismenu et Beaudry, 2002, p. 28). L'objectif principal des prépublications est la communication, contrairement aux publications qui « existe[nt] en raison du processus d'ensemble de sélection, de traitement, de mise en forme, de diffusion, d'institutionnalisation des forums, de reconnaissance et d'archivage des textes soumis » (Boismenu et Beaudry, 2002, p. 29). Les publications, autrement dit, servent à un certain contrôle de la qualité, au processus d'accréditation des professeurs, mais aussi à la conservation des documents.

La proportion d'articles publiés en libre accès est passée, dans l'ensemble des disciplines, de 14% en 1998 à 21% en 2006 et, pour les sciences sociales, de 14% à 38% pour la même période (Gargouri, Larivière, Gingras, Carr et Harnad, 2012). Plus récemment, un important rapport de la firme Science-Metrix a démontré que la proportion d'articles publiés en libre accès avait été sous-estimée dans la littérature et que le « point critique » (le « *tipping point* » qui signifie que plus de 50% des articles sont en libre accès) a déjà été atteint dans plusieurs pays (Archambault et al., 2013). La proportion varie toutefois d'une discipline à l'autre, et les SSH, les champs appliqués, le génie, de même que les technologies présentent la

plus faible proportion d'articles en libre accès. Au moment où le rapport a été publié, le Canada n'avait pas encore passé ce « point critique », mais il est intéressant de noter que le Brésil possédait la plus haute proportion, avec 63% des articles publiés disponibles en libre accès grâce à la base de données Scientific Electronic Library Online (SciELO). Par ailleurs, il a été démontré que plus les mandats institutionnels sont sévères, plus le nombre de publications déposées dans des répertoires augmente (Gargouri et al., 2012) et que les mandats permettent de tripler le taux d'articles en libre accès vert (Gargouri, Larivière, Gingras, Carr et Harnad, 2012). En effet, malgré la sensibilisation et les mesures de soutien mises en place, la proportion d'articles en libre accès demeure entre 10% et 25% en l'absence de mandats, mais elle peut monter à 80% ou plus avec des mandats (Couture, 2013).

Le libre accès en France et au Québec

Suite à une recommandation de la Commission européenne adoptée en juillet 2012, les pays membres de l'Union européenne sont tenus de publier les articles de revues savantes financés par des fonds publics en libre accès six mois après leur publication, ou après douze mois pour les SSH (Commission européenne, 2012, p. 5). À la suite de cette recommandation, le Ministère de l'enseignement supérieur et de la recherche en France a mandaté l'Institut des politiques publiques (IPP) pour rédiger un rapport sur l'impact des embargos sur la diffusion des résultats de recherche en SSH. Publié en juillet 2015, ce rapport a démontré que, dans la plupart des disciplines en SSH, « plus la durée de barrière mobile est longue et plus le nombre de vues annuel de la revue est faible », donc « les revues à barrière mobile courte sont plus vues que les autres » (Bacache-Beauvallet, Benhamou et Bourreau, 2015, p. 55). De plus, « [l]a perte d'audience liée à la barrière mobile apparaît dès un an » (Bacache-Beauvallet, Benhamou et Bourreau, 2015, p. 60).

En réaction aux résultats obtenus par le rapport IPP, une étude, parue en octobre 2015, a été réalisée par IDATE et Cairn.info pour nuancer les conclusions et pallier les lacunes importantes de ce rapport, auquel ils reprochent de s'être « limit[é] à estimer les bénéfices (évidents) qu'apporterait en termes d'audience la mise en place d'un système d'Open Access sans en estimer les coûts [d'implantation] » (IDATE et Cairn.info, 2015a, p. 11). Les auteurs sont arrivés à la conclusion que, le secteur des revues de langue française en SSH étant déjà

fragile en France, la réduction de la durée des embargos, telle que proposée dans le rapport IPP, mettrait en péril la plateforme Cairn.info, mènerait à la disparition de nombreuses revues de référence ou à la diminution de leur qualité et nuirait au rayonnement de la recherche française en SSH à l'international. Pour éviter une « véritable implosion du secteur de l'édition scientifique de langue française (dans le domaine des S.H.S.) », les auteurs proposent l'adoption de mesures d'accompagnement et la transition vers un modèle de libre accès de type « platine », où le gouvernement et les institutions défrayeraient en amont les coûts liés à l'abolition des barrières mobiles, ce qui serait compensé par la diminution des dépenses d'acquisition des bibliothèques universitaires (IDATE et Cairn.info, 2015b).

Au Québec, Boismenu et Beaudry, dès 2002, mettaient en garde contre la diffusion gratuite des revues en ligne sans une réorganisation du financement de la recherche, en affirmant que la gratuité « a[vait] tout lieu d'accentuer la précarité financière des revues et même de leur être fatale » (2002, p. 81). Aujourd'hui, si les revues savantes en libre accès ne semblent pas plus précaires que celles en accès restreint ou différé, les revues francophones au Canada sont, dans l'ensemble, grandement « déstabilisées par le numérique » et « prises dans un vent de désordre, stimulant et déstabilisant » (Lebel, 2016). Les revues estiment également que leur financement devrait provenir des universités et des gouvernements, selon la première phase de l'enquête menée par Éric Duchemin (Lebel, 2016), ce qui aurait tout lieu de réduire leur instabilité financière et de les soutenir dans une transition vers le libre accès. La deuxième phase de l'enquête nous renseignera peut-être sur cette question.

En ce qui concerne les pratiques d'auto-archivage au Québec, une étude menée en 2006 a montré que 27% des chercheurs québécois, à ce moment, avaient déjà publié dans une revue en libre accès, mais seulement 12% des chercheurs avaient auto-archivé l'un de leurs travaux, un faible taux que l'auteur explique par le manque d'information des chercheurs à propos de leurs droits en la matière (Vézina, 2006, p. 6-7, p. 11). Malheureusement, il n'existe que peu de données récentes sur les pratiques des chercheurs au Québec en matière d'auto-archivage. Un sondage réalisé dans le cadre d'un mémoire de maîtrise a montré que, parmi les professeurs de l'ensemble des facultés de l'Université de Sherbrooke, 32% ont déjà publié en libre accès, 53% ont déjà consulté ou cité des articles en libre accès, 68% connaissent le dépôt institutionnel de leur institution et 57% ont un sentiment positif devant l'implantation de ce

dépôt institutionnel (Cloutier, 2015). Néanmoins, parmi ces derniers, 25% n'ont jamais entendu parler du libre accès (Cloutier, 2015). La perception des chercheurs québécois à l'égard du libre accès pourrait changer dans les années à venir avec la nouvelle politique des trois organismes subventionnaires canadiens, soit les Instituts de recherche en santé du Canada (IRSC), le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et le Conseil de recherches en sciences humaines du Canada (CRSH), sur le libre accès aux publications savantes en vigueur depuis le 1er mai 2015 (Gouvernement du Canada, 2016). Cette politique incite les titulaires d'une subvention à diffuser en libre accès leurs résultats de recherche dans les douze mois suivant leur publication.

Impact du libre accès

Plusieurs études tendent à montrer que le libre accès aurait un effet positif sur la diffusion de la science et ce, plus spécifiquement pour les pays en développement. Cependant, l'impact réel du libre accès pris isolément – une fois les biais possibles éliminés – sur le nombre de citations a fait l'objet de vifs débats depuis la parution de l'article de Lawrence dans *Nature* (2001). Ce dernier a obtenu, pour la première fois, une corrélation claire entre la diffusion en libre accès et le nombre de citations obtenues par des actes de conférence dans le domaine de l'informatique (Lawrence, 2001). D'autres études sont arrivées à des conclusions similaires dans la première décennie des années 2000 (p. ex. Antelman, 2004; Eysenbach, 2006; Greyson, Morgan, Hanley et Wahyuni, 2009; Hajjem, Harnad et Gingras, 2005; Harnad et Brody, 2004; Metcalfe, 2005; Metcalfe, 2006; Norris, Oppenheim et Rowland, 2008; Schwarz et Kennicutt Jr, 2004).

Toutefois, à partir de 2005 environ, certains auteurs ont commencé à remettre en cause l'impact positif du libre accès observé dans certaines des études précédentes en leur reprochant d'avoir été biaisées par des variables non contrôlées (Davis, 2011b; Davis et Fromerth, 2007; Davis, Lewenstein, Simon, Booth et Connolly, 2008; Gaulé et Maystre, 2011; Henneken et al., 2006; Kurtz et al., 2005a; Lansingh et Carter, 2009; Moed, 2007). Dans l'une de leurs études, Kurtz et ses collaborateurs ont examiné les causes de la corrélation obtenue par Lawrence et ont suggéré trois explications possibles : l'hypothèse du libre accès (« *Lawrence Effect* » ou « *OA advantage* », selon laquelle les articles seraient cités

plus souvent, car les auteurs y auraient accès plus facilement), l'hypothèse de l'accès rapide (« *early access* » ou « *early view* », selon laquelle les articles seraient cités plus souvent, car ils deviendraient disponibles plus rapidement) et l'hypothèse du biais d'auto-sélection (« *self-selection bias* », selon laquelle les articles seraient cités plus souvent, car les auteurs préfèreraient rendre leurs articles les plus importants – donc les plus susceptibles d'être cités – en libre accès) (2005a, p. 2). Ils sont parvenus à la conclusion qu'il y a un important effet de l'accès rapide, de même qu'un important biais d'auto-sélection, mais aucune indication d'un « avantage du libre accès » pris isolément (Kurtz et al., 2005a, p. 11). Parmi les études remettant en doute l'« *OA advantage* », notons également l'essai randomisé contrôlé de Davis et ses collaborateurs qui ont cherché à contrôler le « *self-selection postulate* » en diffusant au hasard en libre accès certains articles publiés dans des revues hybrides (Davis, Lewenstein, Simon, Booth et Connolly, 2008). Ils ont conclu que le libre accès n'avait aucun effet sur le nombre de citations au cours de l'année suivant la publication, mais cette étude a été vivement critiquée, entre autres, à cause de la période d'observation trop courte (Eysenbach, Gunther, 2008; Harnad, 2008).

Depuis 2010 environ, plusieurs études ont apporté de nouveaux arguments dans le débat et ont montré un impact positif du libre accès sur la diffusion des résultats de recherche (p. ex. Archambault et al., 2013; Björk et al., 2010; Gargouri et al., 2010; Gargouri et al., 2012; Gargouri, Larivière, Gingras, Carr et Harnad, 2012; Gentil-Beccot, Mele et Brooks, 2010; Lippi et Favaloro, 2012; Mounce, 2015; Riera et Aibar, 2013; Wang, Liu, Mao et Fang, 2015). Certains auteurs, par ailleurs, présentent des résultats mitigés, tels que Calver et Bradley (2010) (un impact positif du libre accès sur le nombre de citations a été observé pour les chapitres de livres, mais pas pour les articles), de même que McCabe et Snyder (2011) (un impact positif a été observé sur la plateforme JSTOR, mais pas sur ScienceDirect). Parmi les études confirmant un effet positif du libre accès, Gargouri et ses collaborateurs ont démontré qu'il n'y a aucun « biais d'auto-sélection » qui puisse amplifier l'effet du libre accès :

> the OA advantage is a statistically significant, independent positive increase in citations, even when we control the independent contributions of many other salient variables (article age, journal impact factor, number of authors, number of pages, number of references cited, Review, Science, USA author) (2010, p. 8).

Les auteurs ne choisiraient donc pas de diffuser gratuitement en ligne leurs « meilleurs » travaux ; l'« *OA advantage* » proviendrait plutôt du fait que les lecteurs peuvent choisir de citer les travaux de plus grande qualité parmi une large sélection, sans être limités aux revues auxquelles leur institution est abonnée (Gargouri et al., 2010). Le rapport de Science-Metrix (Archambault et al., 2013) a également démontré l'impact positif du libre accès sur les citations pour les articles publiés en libre accès vert et hybride dans toutes les disciplines étudiées, en particulier celles qui présentent un proportion d'articles publiés en libre accès plus faible, soit les SSH, les champs appliqués, le génie et les technologies. Néanmoins, les auteurs ont observé un impact négatif pour le libre accès doré, car les revues dorées sont souvent récentes et plus petites. Enfin, Wang, Liu, Mao et Fang (2015) ont constaté un impact positif du libre accès non seulement sur le nombre de citations (de 1,11 à 1,52 fois plus élevé pour les articles en libre accès), mais également sur le nombre de vues (de 2,5 à 3,6 fois supérieur pour les articles en libre accès) et les partages sur Facebook et Twitter (de 1,47 à 1,22 fois plus élevé pour les articles en libre accès). Ils ont également observé que l'écart entre les articles en libre accès et ceux en accès restreint s'agrandit avec le temps[55].

Pays en développement

Bien que la corrélation entre le libre accès et le nombre de citations ne fasse pas toujours pas l'unanimité, les auteurs s'entendent généralement pour dire aujourd'hui que le libre accès a un impact positif sur la diffusion des résultats de recherche et cela, encore plus pour les pays en développement. Gaulé (2009) a remarqué, d'une part, que les listes de références des chercheurs indiens est 8,9% plus courte en biologie et 10,8% plus courte en médecine que celles des chercheurs suisses qui publient dans les mêmes revues, ce qui suggère un important problème d'accès à la littérature scientifique, et d'autre part, que les listes de références des chercheurs indiens contiennent 50% plus de références à des revues en libre accès que celles des chercheurs suisses. Evans et Reimer (2009) ont, de leur côté, démontré que le libre accès a un effet positif deux fois plus important dans les pays en développement de l'hémisphère sud que dans les pays développés des hémisphères nord et

[55] Signalons toutefois les limites de cette étude : la petite taille de l'échantillon, l'asymétrie entre le nombre d'articles en libre accès et en accès restreint, l'utilisation d'une seule revue et la courte période de temps observée.

ouest, sauf dans les pays les plus pauvres où l'accès à Internet est limité et où, par le fait même, l'impact du libre accès est moins apparent. Malgré cela, une autre étude a conclu que les auteurs provenant de pays en développement dans le domaine de la biologie ne publient pas davantage dans des revues en libre accès ni ne citent davantage d'articles en libre accès (Frandsen, 2009). On reproche toutefois à Frandsen la petite taille de son échantillon (Riera et Aibar, 2013, p. 8) et, selon Gaulé, l'analyse menée ne permet pas d'exclure statistiquement un effet positif du libre accès (2009, p. 3). Enfin, Davis a constaté que si l'accès à The Essential Electronic Agricultural Library (TEEAL), disponible sans connexion à Internet et à très faible coût pour les pays en développement, ne mène pas à une plus grande production d'articles scientifiques en Afrique subsaharienne, il a pour effet d'allonger les listes de références (2,6 références supplémentaires par publication) et un plus grand nombre de citations à des revues de cette base de données (Davis, 2011a).

Voici qui conclut notre revue de littérature. Ce premier chapitre met en lumière la contribution de notre projet de recherche à différentes questions trop peu abordées jusqu'à présent. Si l'évolution de la situation des revues savantes en SSH au Canada et au Québec, ainsi que l'évolution des pratiques d'auto-archivage des chercheurs québécois n'ont pas été décrites depuis plus de dix ans, notre étude permet de mieux comprendre l'usage des revues savantes au Québec et l'impact des politiques de libre accès des revues sur leur usage. En outre, devant le manque constaté dans le champ de la webométrie, nous contribuons aux études basées sur l'analyse des données de téléchargement. Ce manque est d'ailleurs encore plus important pour les SSH que pour les sciences naturelles. Dans cette revue de littérature, nous avons abordé les principales questions qui éclaireront notre analyse de l'usage des articles de revues savantes sur la plateforme Érudit, nommément les particularités de la communication savante en SSH, l'obsolescence de la littérature savante, la webométrie, l'analyse des données de téléchargements, ainsi que le libre accès. Le chapitre suivant décrit notre méthodologie.

Chapitre 2 : Méthodologie

Notre projet de recherche est de type descriptif quantitatif – un type qui, selon Fortin et Gagnon, permet de décrire « de[s] concepts, de[s] facteurs ou de[s] caractéristiques propres à une population » (2010, p. 125). Dans notre cas, nous cherchons à décrire les caractéristiques propres à l'usage de la plateforme Érudit, autant du point de vue des articles téléchargés que du point de vue des utilisateurs. Tel que présenté dans l'introduction, nos résultats sont organisés en deux parties : la première constituant une analyse exploratoire de l'usage de la plateforme Érudit et la seconde, s'intéressant au libre accès, comportant un aspect quasi-expérimental avec la comparaison de la courbe d'usage des articles savants avant et après la fin de la période d'embargo, de même qu'avec la comparaison entre les revues en accès différé et celles en libre accès complet. Étant donné la variété des questions de recherche qui sont explorées, notre méthodologie est basée sur différentes études portant sur l'obsolescence de la littérature scientifique et sur le libre accès, mais plus particulièrement sur des études utilisant l'indicateur webométrique des données de téléchargements.

Source des données

Contexte

La plateforme Érudit a connu sa première phase de développement en 1998 grâce à une subvention du Fonds FCAR, puis, en 2004, le Consortium Érudit a été créé pour regrouper l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal en une même société (Beaudry, Boucher, Niemann et Boismenu, 2009). De nos jours, la collection de la plateforme rassemble environ 200 000 documents, dont ceux publiés par plus de 150 revues savantes et culturelles, ainsi que des livres, des actes de conférence, des mémoires et thèses, des rapports et des notes de recherche. Érudit jouit d'une grande visibilité à l'international, malgré sa couverture essentiellement en SSH – une couverture, par définition, très ancrée géographiquement. En effet, ses abonnés sont répartis dans plus de 35 pays et les utilisateurs proviennent de plus de 200 pays (Érudit, 2015). La grande majorité des ressources sur Érudit (environ 95%) est en libre accès (Érudit, s.d.-b). Cette plateforme combine différents modèles

d'accès : le libre accès doré (20 revues savantes en libre accès complet, c'est-à-dire sans période d'embargo), le libre accès différé (pour rappel, il s'agit d'une forme d'accès doré, mais avec une période d'embargo), mais également le libre accès vert (Érudit diffuse des publications « dans un dépôt à 37 centres de recherche et universités canadiennes » [Érudit, s.d.-b]). À l'origine, les revues savantes étaient entièrement diffusées en libre accès sur la plateforme, mais une barrière mobile a été instaurée en 2006 pour la majorité des revues, à la demande des directions de revues (Beaudry, Boucher, Niemann et Boismenu, 2009).

Disciplines couvertes

Dans cette section, nous dresserons un portrait des disciplines couvertes dans la base de données d'Érudit en utilisant la classification du National Science Foundation (NSF) – classification utilisée dans l'ensemble de ce mémoire. Malgré les imperfections que l'on peut lui reprocher, nous avons préféré cette classification à celle employée sur le site Web d'Érudit, car il s'agit d'un schéma hiérarchique utilisé internationalement. Basé sur trois niveaux, soit les grandes disciplines (p. ex. : sciences sociales et humaines), les disciplines (p. ex. : humanités) et les spécialités (p. ex. : littérature), ce schéma permet de regrouper les téléchargements de différentes manières pour raffiner notre analyse. Le schéma de classification d'Érudit, au contraire, n'est pas hiérarchique et plusieurs disciplines peuvent être attribuées à une même revue (jusqu'à trois), ce qui complexifie grandement le comptage des articles d'une discipline. Nous ne l'avons donc pas retenu pour ce projet.

Sur les 106 revues savantes de la plateforme Érudit qui font l'objet de notre étude[56], 99 appartiennent à la grande discipline des sciences sociales et humaines (SSH) (93,4% des revues), contre sept seulement en sciences naturelles et génie (SNG) (6,6% des revues) (voir l'annexe 1). Une importante proportion des revues en SSH (38,4% du nombre de revues) font partie de l'une ou l'autre spécialité des humanités, soit, en ordre décroissant d'importance : l'histoire, divers champs des humanités (c.-à-d. des revues multidisciplinaires essentiellement, mais pas exclusivement, en sciences humaines, ainsi que des revues associées à un groupe ou une culture, p. ex. études anglaises, africaines, québécoises, etc.), la littérature (à noter que la

[56] Rappelons que la plateforme rassemblait 119 revues savantes (si l'on agrège tous les titres possibles d'une même revue) au moment de la collecte de données, mais que les treize revues du fonds UNB (Érudit, 2016) ne sont pas enregistrées dans les logs de serveurs ni dans les données d'articles d'Érudit.

spécialité « littérature » inclut les études cinématographiques dans la classification du NSF), les langues et la linguistique, la religion et la philosophie. Viennent ensuite les revues de la discipline des sciences sociales, qui représentent 31,3% des revues en SSH, dont les spécialités sont les suivantes, en ordre décroissant d'importance : la sociologie, l'anthropologie et l'archéologie (ce qui inclut l'ethnologie), l'économie, la géographie, les sciences politiques et l'administration publique, les sciences sociales générales (c.-à-d. des revues multidisciplinaires essentiellement, mais pas exclusivement, en sciences sociales), diverses sciences sociales (c.-à-d. des revues dans des domaines considérés « autres » par le NSF comme l'éthique ou les études féministes), les relations internationales (incluant les études militaires), la criminologie, la démographie et l'étude sur la science. Les champs professionnels sont la discipline qui occupe la troisième place en termes de nombre de revues dans la grande discipline des SSH (avec 21,2% des revues dans cette discipline). Les spécialités des champs professionnels sont les suivantes : l'éducation, le travail social, le droit, le management, la bibliothéconomie et l'archivistique, ainsi que divers autres champs professionnels (tels que le tourisme). On trouve, enfin, quelques revues dans d'autres spécialités, soit les arts du spectacle (ce qui inclut la musique et le théâtre), les beaux-arts et l'architecture, la psychanalyse, la gériatrie et la gérontologie (comprenant les études sur la mort) et la réhabilitation (incluant les études sur la déficience intellectuelle). Enfin, parmi les revues qui appartiennent aux SNG, trois sont dans la spécialité science environnementale, deux en psychiatrie, une en botanique et une en recherche biomédicale générale.

Données de logs

L'ensemble des données de téléchargement des articles de revues savantes d'Érudit couvre la période allant du 1er avril 2010 au 31 décembre 2015. Les données sont regroupées dans des fichiers texte ; chaque fichier texte correspondant aux téléchargements d'une journée. Elles contiennent les informations suivantes : la date et l'heure à laquelle le téléchargement a eu lieu (selon l'heure normale de l'Est) ; l'adresse IP de l'usager ; l'adresse IP du proxy utilisé par l'usager (s'il y a lieu) ; l'abréviation du titre de la revue ; l'année de publication de l'article ; le volume ; le numéro ; l'identifiant unique de l'article téléchargé (constitué de six chiffres, des lettres « ar » et de l'extension .pdf ou .html) ; l'URL à partir de laquelle l'usager a accédé à l'article (s'il y a lieu) ; le navigateur utilisé par l'usager ; la version de son système

d'exploitation[57]. Les logs d'Érudit correspondent donc globalement au Apache Common Log Format, standard pour les données d'usage Web (Kurtz et Bollen, 2010, p. 10). Bien que ces données soient bien structurées, un travail important de nettoyage a dû être effectué pour isoler les téléchargements d'articles de revues savantes, avec un effort particulier en ce qui concerne la détection de robots.

Particularités de la plateforme Érudit

Certaines particularités de la plateforme ont dû être prises en compte dans la réalisation de cette étude. Elles ont influencé le traitement et l'analyse des données, et certaines sont liées aux limites présentées plus loin.

1. Il n'existe aucun fichier de logs avant la date du 1er avril 2010. Selon l'équipe technique d'Érudit, aucun des employés présents à l'époque ne se souvient si les logs étaient enregistrés auparavant ou non et, si c'était le cas, s'ils ont été perdus. Les logs sont désormais conservés de façon ininterrompue (à l'exception de quelques journées où le serveur a dû être redémarré), mais nous avons choisi de nous arrêter en date du 31 décembre 2015. Lors de l'analyse des données, nous avons parfois dû nous restreindre à la période 2011-2015 pour couvrir cinq années complètes. Cela est précisé dans les résultats.

2. Le passage au libre accès s'effectue au même moment, en janvier, pour tous les numéros d'une année. Par exemple, en janvier 2016, tous les numéros publiés au cours de l'année 2014 sont passés au libre accès en même temps, y compris ceux publiés en décembre 2014. Par conséquent, la période d'embargo est d'un maximum de 24 mois, mais elle est souvent plus courte.

3. Le titre de certaines revues a changé au cours de leur histoire et, parfois même, à plus d'une reprise[58]. L'abréviation du titre des revues, qui est employée dans l'URL du site d'Érudit et qui nous permet d'identifier les articles téléchargés dans les fichiers de logs, est modifiée

[57] Voici un exemple de données brutes : 2010-04-01 00:00:11 132.204.2.131 GET /revue/nps/2005/v17/n2/011223ar.pdf HTTP/1.1 - 80 - 132.208.105.77 "Mozilla/5.0 (Windows; U; Windows NT 5.1; fr; rv:1.9.2.2) Gecko/20100316 Firefox/3.6.2" "http://www.erudit.org/revue/nps/2005/v17/n2/" 200 343622.

[58] Les revues peuvent connaître différentes transformations. Elles peuvent notamment fusionner, être vendues à un éditeur, changer de titre, et plus encore (Library of Congress, 1994). Cependant, dans notre cas, il n'y a eu que des changements de titre.

lorsqu'une revue change de titre. Il est donc essentiel d'associer tous les titres possibles d'une revue pour éviter de les considérer comme des revues différentes.

4. Plusieurs revues ont pris du retard dans leur publication. Si l'année courante est 2016, par exemple, il est possible que le dernier numéro mis en ligne soit 2013 pour une revue donnée. La durée de l'embargo devrait idéalement être calculée à partir de la date de mise en ligne des numéros de revue sur le site de la plateforme, et non pas à partir de la date de publication donnée par la revue. Toutefois, malgré de nombreux efforts de la part de l'équipe technique d'Érudit, il n'a pas été possible d'identifier avec certitude la date de mise en ligne des numéros qui nous aurait permis de calculer avec précision la date de leur passage au libre accès. Par conséquent, pour les articles publiés depuis le 1er avril 2010 – date de début des données de logs –, nous avons considéré la date du premier téléchargement d'un article dans un numéro donné comme la date de mise en ligne de l'ensemble des articles de ce numéro et, pour les articles publiés avant le 1er avril 2010, nous avons utilisé la date de publication donnée par la revue[59]. En raison de ce biais important, nos résultats portant sur l'effet du passage au libre accès se limitent aux articles publiés depuis 2011 (de manière à couvrir cinq années complètes).

5. Une version PDF du texte intégral est disponible pour tous les articles sur la plateforme Érudit (c'est ce que l'équipe d'Érudit appelle le traitement minimal). En plus du format PDF, le texte intégral est parfois disponible en format HTML (c'est-à-dire le traitement complet)[60]. Le format de diffusion, de manière générale, est le même pour tous les articles d'un même numéro, mais, selon les informations que nous avons colligées, 0,12% des articles qui n'ont qu'un format PDF se trouvent dans un numéro où la majorité des articles ont à la fois une version PDF et une version HTML, tandis que 0,04% des articles qui sont en format PDF et HTML se trouvent dans un numéro où la majorité des articles n'ont qu'une version PDF. Lors du nettoyage des données, il est essentiel de retirer les téléchargements dont l'URL se termine

[59] Il n'y a aucun précédent dans la littérature de l'utilisation de cette technique, car le problème rencontré est très spécifique à notre jeu de données. Cette technique a été développée en collaboration avec Yorrick Jansen, ingénieur en informatique, et nous pensons qu'elle est fiable dans la très grande majorité des cas. À notre avis, les cas où aucun article d'un numéro n'est téléchargé pendant un an doivent être isolés et ils n'influencent sans doute pas les tendances observées.

[60] Nous remercions l'équipe technique d'Érudit de nous avoir fourni les informations détaillées concernant le format de diffusion de l'ensemble des numéros de revues de la plateforme.

par « .html » pour les articles qui n'ont qu'une version PDF (en traitement minimal), car ils ne correspondent pas au téléchargement du texte intégral, mais seulement à une vue des métadonnées de l'article. On conserve toutefois, pour les articles en traitement complet, les téléchargements dont l'URL se termine par « .html » ou « .html?vue=integral ». Par souci de commodité – et étant donné le très faible taux d'erreur que cela introduit –, nous avons considéré que l'ensemble des articles d'un numéro possède le format de diffusion de la majorité (autrement dit, nous avons arrondi au numéro près). En outre, nous avons arrondi à l'année près plutôt qu'au volume, qu'au numéro ou qu'à l'article près ; toutefois, nous avons été le plus restrictifs possible, c'est-à-dire que dans les cas où une partie seulement des numéros d'une année sont en traitement complets, nous avons considérés que cette année est en traitement minimal. Cela permet de limiter la surreprésentation des articles en traitement complet[61].

6. Par ailleurs, pour les articles qui n'ont qu'un format PDF, aucun traitement n'est fait *a posteriori*, après la mise en ligne des articles, pour produire une version HTML (ou alors, quelques cas très marginaux seulement). Ceci signifie qu'un usager ne peut pas consulter un article qui ne possédait, au moment de sa mise en ligne, qu'un format PDF, puis le consulter de nouveau deux ans plus tard, par exemple, en format HTML. Cependant, tous les numéros d'une même revue n'ont pas nécessairement le même format de diffusion ; en effet, nous avons relevé 20 cas où une revue qui était passée au traitement complet est revenue au traitement minimal pour certains numéros (dont six revues où cela s'est produit plus d'une fois). Nous avons donc tenu compte du format de diffusion de chaque numéro plutôt que de considérer qu'une revue est passée au traitement complet à partir d'une date donnée.

Afin de tenir compte des particularités ci-dessus au moment du traitement des données, nous avons développé un référentiel de revues, qui respecte le format de données textuelles JSON (c.-à-d. *JavaScript object notation*). Nous y avons rassemblé les métadonnées des

[61] De surcroît, cela permet de limiter le nombre de doublons : lorsqu'un utilisateur consulte la page Web d'un article sur Érudit, qu'il soit en traitement minimal ou complet, puis qu'il télécharge la version PDF, cela compte comme une vue et un téléchargement (dans le cas du traitement minimal) ou comme deux téléchargements distincts (dans le cas du traitement complet).

revues provenant de différents rapports fournis par l'équipe technique d'Érudit, mais nous avons procédé nous-mêmes à la classification des revues en nous conformant au schéma NSF. Pour ce faire, nous avons d'abord tenu compte de la classification utilisée sur le site d'Érudit, puis nous avons survolé le contenu de chaque revue pour nous assurer de trouver la meilleure correspondance possible dans la classification du NSF. La figure 2 montre, en guise d'exemple, les métadonnées d'une revue fictive suivant la syntaxe de notre référentiel de revues :

```
[
  {
    "id": "abcd",
    "other_ids": ["ab12", "abc123"],
    "names": [
      {"url_name": "ab12", "full_name": "Revue internationale AB", "start_year": 1956, "stop_year": 1965},
      {"url_name": "abc123", "full_name": "Revue internationale ABC", "start_year": 1966, "stop_year": 1989},
      {"url_name": "abcd", "full_name": "Revue internationale ABCD", "start_year": 1990}
    ],
    "full_text_html": [
      {"start_year": 2003, "stop_year": 2007},
      {"start_year": 2009}
    ],
    "general_discipline_fr": "Sciences sociales et humaines",
    "general_discipline": "Social Sciences and Humanities",
    "discipline_fr": "Humanités",
    "discipline": "Humanities",
    "speciality_fr": "Histoire",
    "speciality": "History",
    "full_oa": false
  } ]
```

Figure 2. Référentiel de revues

Les données relatives à chacune des revues savantes d'Érudit sont enregistrées de cette manière dans le référentiel de revues. Le « id » est l'identifiant unique attribué à chaque revue pour tous les titres qu'elle a pu avoir à travers le temps. Nous avons utilisé l'« url_name » le plus récent en guise d'identifiant unique. Sous « names », on trouve la liste de tous les titres possibles d'une revue. Plus spécifiquement, pour chaque titre qu'a eu la revue dans le temps, on trouve l'abréviation utilisée dans l'URL (« url_name »), le titre complet de la revue (« full_name »), ainsi que l'intervalle de temps pendant lequel ce titre a été utilisé (« start_year » et « stop_year »)[62]. Sous « full_text_html », on mentionne, s'il y a lieu, le ou les intervalles de dates pendant lesquels les numéros de revue ont été diffusés en traitement complet (c.-à-d. qu'ils ont une version HTML du texte intégral en plus de la version PDF). Lorsque les crochets carrés sont vides (« "full_text_html": [ ] »), cela signifie que la revue n'a

[62] Notons que les dates utilisées correspondent aux dates de publication données par la revue, non pas aux dates de mise en ligne sur le site d'Érudit.

jamais été diffusée en traitement complet. L'absence de « stop_year » pour le titre « Revue internationale ABCD », à la fois dans le cas des titres et dans le cas du format de diffusion, signifie que ce titre est encore utilisé aujourd'hui, ou encore, si la revue a cessé de paraître, que ce titre a été utilisé jusqu'à la fin de son existence. Viennent ensuite les disciplines en français et en anglais selon la classification du NSF, de la discipline générale à la spécialité. Enfin, « full_oa » permet de savoir si la revue est en libre accès complet (« true ») ou en libre accès différé (« false »).

Traitement des données

Étapes suivies pour le nettoyage des données

En ce qui concerne le nettoyage des données brutes, voici les étapes que nous avons suivies. Un pré-filtre n'a conservé que les lignes :

- qui contiennent une requête http « GET », permettant de récupérer une ressource sur une page Web ;
- dont le code de retour est 200, ce qui indique que la requête a bien fonctionné ;
- dont la requête débute par « /revue », afin de nous restreindre aux revues savantes ;
- et dont la ressource téléchargée est un article qui se termine par « ar.pdf », « ar.html?vue=integral », ou encore « ar.html » pour les articles qui ont une version HTML en plus de la version PDF.

Les activités autres que les téléchargements d'articles savants (p. ex. les téléchargements de JavaScripts, de CSS, d'images, etc.) ont donc été exclues, de manière à réduire le temps d'exécution du script. Ensuite, nous avons éliminé les quelques lignes mal formatées dont les informations ne pouvaient être extraites. Enfin, l'étape la plus complexe a été de distinguer les téléchargements d'articles par des robots de ceux par des humains, les seuls qui nous intéressent. Les robots connus (par « connus », nous entendons les robots qui se déclarent comme tels, par exemple ceux qui indexent les sites Web) ont été facilement retirés grâce à l'agent utilisateur (au moyen d'une liste standardisée de robots). Le tableau 1 illustre la proportion de données intéressantes pour notre étude, ainsi que la réduction de la taille des fichiers après le nettoyage.

Tableau 1. Nettoyage des données de logs

Nombre de fichiers de logs	2 062
Nombre total de requêtes HTTP dans les fichiers de logs	999 367 190
Pourcentage de requêtes HTTP analysables dans les fichiers de logs	99,99%
Pourcentage de requêtes HTTP qui réfère à des articles savants (humains et robots)	10,34%
Pourcentage de requêtes HTTP qui réfère à des articles savants (humains seulement)	3,95%
Nombre total de téléchargements d'articles savants par des humains	39 437 659

Suite à ce pré-filtre, les revues qui ont changé de titre ont été regroupées sous un identifiant unique et les revues en libre accès complet ont été clairement identifiées (afin d'éviter de les prendre en compte lors de l'analyse de l'effet de la fin de l'embargo). Les coordonnées géographiques indiquant le lieu à partir duquel l'article a été téléchargé ont été ajoutées dans les fichiers CSV, de même que le fuseau horaire[63]. Ceci nous a permis d'ajouter l'heure locale du téléchargement – mais nous n'avons pas tenu compte, pour des raisons techniques, des changements d'heure entre l'heure d'été et l'heure d'hiver dans les différents pays qui ont téléchargé des articles sur Érudit. En outre, nous avons enlevé les caractères spéciaux et uniformisé les champs contenant le navigateur et le système d'exploitation utilisés par les usagers[64]. L'âge des articles téléchargés a été ajouté en soustrayant l'année de publication de l'année du téléchargement[65]. L'annexe 2 présente les différentes informations contenues dans les logs, ainsi que celles qui ont été calculées et ajoutées aux logs au moment du traitement des données.

Le script de nettoyage de données est prévu pour s'exécuter depuis un disque dur externe qui contient l'ensemble des fichiers de logs, de manière à faire le nettoyage en amont, avant de mettre les données sur un serveur. Une base de données relationnelle, qui contient les informations relatives aux téléchargements d'articles de revues savantes par des êtres humains,

[63] La librairie Python GeoIP 1.2 nous a permis d'accéder aux bases de données GeoIP. Ces dernières associent à une adresse IP plusieurs renseignements, dont le continent, le pays, la ville et le fuseau horaire.

[64] Ces informations ont été identifiées grâce à la librairie Python User-Agents 1.0.1 qui analyse la chaîne de caractères de l'agent utilisateur.

[65] Notons que notre calcul de l'âge des articles est arrondi à l'année près, comme le sont les fichiers de logs qui ne donnent que l'année de publication, sans le mois. Ceci a pour effet d'introduire un biais (négligeable), puisque les articles publiés plus tôt dans l'année ont eu davantage de temps pour accumuler des téléchargements que les autres. Pour les articles publiés depuis le 1er avril 2010, le calcul de l'âge est basé sur la date du premier téléchargement d'un article dans un numéro donné et, pour les articles publiés avant le 1er avril 2010, le calcul est basé sur la date de publication donnée par la revue. Dans un projet futur, il serait toutefois possible de croiser les fichiers de logs avec les données des articles d'Érudit pour obtenir plus de précisions dans le calcul de l'âge.

les métadonnées des revues et, enfin, certaines métadonnées des articles de la collection d'Érudit, a été conçue. L'annexe 3 illustre la structure de cette base de données relationnelle.

Robots

En observant les données, nous avons pris conscience de la présence non négligeable de robots qui cherchent à se faire passer pour des individus (pour rappel, les robots connus ont été éliminés en amont au moyen d'une liste standardisée). Au cours d'une première exploration des données de téléchargements, nous avons constaté avec surprise un biais considérable introduit par les robots malicieux. Dans la figure 3, le pic que l'on observe est causé par une seule adresse IP qui a téléchargé 13 656 fois le même article de 1968 sur Érudit en date du 6 février 2013. Cette figure est basée sur un échantillon non probabiliste de cinq fichiers de logs, soit les premiers mercredis de chaque année, le 2 février 2011, le 1er février 2012, le 6 février 2013, le 5 février 2014 et le 4 février 2015. La première version du script, utilisée dans cette figure, détermine un seuil de téléchargements par jour à partir duquel on considère l'activité d'un utilisateur comme suspecte ; après quelques essais, nous avons opté pour un seuil de plus de 300 téléchargements d'articles savants par jour. Les adresses IP des utilisateurs ont été vérifiées manuellement à l'aide d'un outil de géolocalisation d'adresses IP (IP Location, 2016) pour exclure les robots, mais conserver les proxys d'institutions universitaires. Afin de comparer la courbe avec et sans les robots malicieux, le script a été conçu de manière à pouvoir être exécuté en conservant ou en excluant les robots. Même si, à ce stade, notre échantillon était trop petit pour être représentatif de l'ensemble, le pic observé nous a convaincus de mettre en place une technique de détection de robots plus sophistiquée qui sera décrite plus loin pour éviter de telles anomalies.

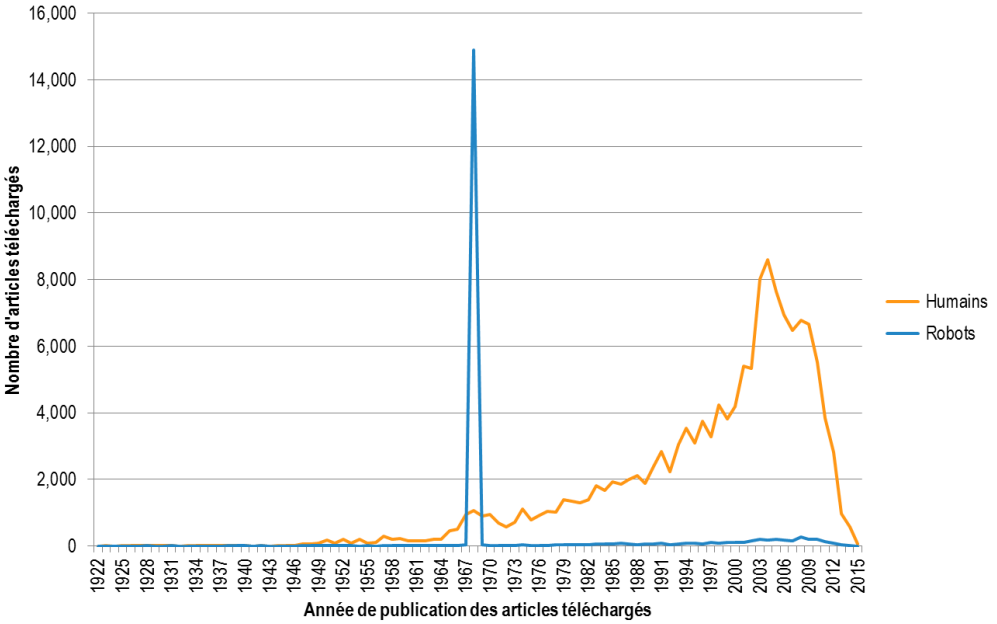

Figure 3. Téléchargements par année de publication (robots et humains) pour les journées sélectionnées 2011-2015 (n = 164 057)

Les robots peuvent être définis comme des « systèmes autonomes qui envoient des requêtes à des serveurs pour obtenir des ressources sur le Web »[66] (Doran et Gokhale, 2010, p. 184). Parmi les bons robots, on trouve les « *crawlers* » qui parcourent le Web dans le but d'indexer le contenu des pages Web. Les moteurs de recherche tels que Google sont basés sur l'utilisation de tels robots. Les mauvais robots, quant à eux, peuvent être simplement des robots mal conçus dont le comportement n'est pas celui initialement espéré. Ils peuvent également avoir des intentions malveillantes, comme créer des pages imitant un site Web connu, rechercher des adresses électroniques pour l'envoi de pourriels ou parcourir des sites Web afin d'y trouver de failles de sécurité.

Doran et Gokhale (2010) décrivent quatre catégories de techniques de détection de robots, allant de la plus élémentaire à la plus efficace. Les trois premières catégories sont des techniques de détection hors ligne – les seules qui peuvent s'appliquer aux logs de serveurs –, tandis que les techniques de la dernière catégorie agissent en temps réel – ce sont les plus efficaces, mais ce sont également les plus coûteuses et les plus difficiles à mettre en place (voir l'annexe 4). Le premier type de détection hors-ligne est l'analyse syntaxique des logs[67] (Doran et Gokhale, 2010, p. 189-191). Il s'agit simplement de rechercher des mots-clés dans le champ « agent utilisateur » ou de rechercher les adresses IP listées comme étant des robots. Le principal problème de cette technique est qu'elle ne repêche que les robots connus ; or, de nouveaux robots qui cherchent à passer inaperçu sont créés quotidiennement et il est impossible de garder les listes à jour. Le second type est l'analyse des caractéristiques du trafic[68] (Doran et Gokhale, 2010, p. 191-196). Il permet de repérer les différences dans le comportement des robots et celui des êtres humains sur le Web en observant, par exemple, l'activité au fil des heures de la journée, la durée des sessions, le type de ressources demandées, le nombre de requêtes, le pourcentage d'erreurs, etc. Cette catégorie de techniques est basée sur le comportement attendu des êtres humains et le seuil à partir duquel on juge un comportement anormal. Le troisième type de techniques se fonde sur l'apprentissage

[66] « [A]utonomous systems that send requests to Web servers across the Internet to request resources » (Doran et Gokhale, 2010, p. 184).
[67] « [S]yntactical log analysis » (Doran et Gokhale, 2010, p. 189-191).
[68] « [T]raffic pattern analysis » (Doran et Gokhale, 2010, p. 191-196).

automatique (« *machine learning* ») ou sur des modèles probabilistes[69] (Doran et Gokhale, 2010, p. 196-200). Très efficaces, ces techniques peuvent utiliser des indicateurs variés (p. ex. temps entre les requêtes, durée des sessions, ratio entre les requêtes html et les requêtes d'images, le pourcentage de requêtes HEAD, etc.) pour raffiner leur algorithme, mais elles doivent utiliser un jeu de données d'apprentissage où les robots sont suffisamment nombreux pour garantir une bonne précision dans la détection. Enfin, la dernière catégorie est similaire au test de Turing qui permet de classifier, en temps réel, une conversation produite par un humain ou par une machine[70] (Doran et Gokhale, 2010, p. 200-203). Dans ce cas-ci, la conversation correspond à des paires composées d'une requête http et d'une réponse. Un test (p. ex. un test CAPTCHA) est donc présenté à toutes les sessions actives et la réponse est analysée pour déterminer s'il s'agit d'un robot ou d'un être humain. Les auteurs parviennent à la conclusion que la troisième catégorie est sans doute celle qui présente le plus grand potentiel pour l'avenir de la détection de robots.

Dans le cadre de ce projet, nous avons opté pour une technique appartenant à la deuxième catégorie. Une étude menée en 2006 par Geens, Huysmans et Vanthienen, qui combine différentes techniques d'analyse syntaxique et d'analyse des caractéristiques du trafic (deux premières catégories), propose une règle pour la détection des robots malicieux (2006, p. 128) synthétisée de la sorte par Doran et Gokhale :

> *if a session contains a request for robots.txt OR its IP address is in the robot IP list OR its user-agent field is in the robot user-agent list OR the HEAD method is used OR (the referring field is unassigned AND no images are requested) then classify the session as a Web robot* (2010, p. 193).

Testée sur un jeu de données de logs contenant 241 robots, cette règle obtient un taux de rappel de 97,51% et de précision de 89,35% (Geens, Huysmans et Vanthienen, 2006, p. 128)[71]. Avec un tel taux de précision, un certain nombre de faux positifs sont repêchés ; autrement dit, 10,65% du trafic identifié comme des robots sont, en fait, des humains. Cependant, la précision est tout de même assez élevée pour ne pas fausser les tendances sur un

[69] « [A]nalytical learning techniques » (Doran et Gokhale, 2010, p. 196-200).
[70] « Turing test systems » (Doran et Gokhale, 2010, p. 200-203).
[71] Dans le cas précis de la détection de robots, le rappel est le nombre de robots identifiés divisé par le nombre réel de robots, tandis que la précision est le nombre de robots identifiés divisé par le nombre de robots prévus (Geens, Huysmans et Vanthienen, 2006, p. 125).

gros volume de données. Nous avons adopté cette technique de détection de robots pour notre étude. Néanmoins, après quelques essais sur notre jeu de données, nous l'avons adaptée et raffinée comme suit :

SI (le référent est vide ET aucune image n'est téléchargée ET l'adresse IP télécharge plus de 100 articles savants / jour)

OU (aucun JavaScript n'est téléchargé ET aucune feuille CSS n'est téléchargée ET l'adresse IP télécharge plus de 100 articles savants / jour)

ALORS considère l'adresse IP comme un robot.

Le seuil de téléchargements par jour à partir duquel on considère l'activité d'un utilisateur comme suspecte a été abaissé à 100 de manière à obtenir plus de précision dans l'identification des robots qui téléchargent un faible nombre d'articles. Suite à la mise en place de cette technique de détection de robots, nous avons exécuté le script sur un échantillon aléatoire stratifié (par année de 2010 à 2015) d'une taille de trente journées de logs avec et sans les robots ; pour des raisons d'espace disponible et de temps d'exécution du script, il ne nous a pas été possible d'utiliser un échantillon d'une plus grande taille en conservant les robots. Les résultats observés avec un échantillon aléatoire de 30 journées étaient similaires à ceux de l'échantillon non probabiliste de cinq journées (figure 3), hormis l'imposant pic du 6 février 2013 – une journée qui ne s'est pas retrouvée dans l'échantillon aléatoire subséquent. Les robots n'ont, le plus souvent, qu'un faible impact sur la tendance générale des courbes observées (c'est le cas notamment pour l'âge des articles téléchargés, le nombre de téléchargements par pays, le nombre de téléchargements par référent et le nombre d'articles en libre accès et en accès restreint téléchargés), mais un impact plus important sur le nombre de téléchargements par heure de la journée notamment.

Limites de l'étude

L'analyse de données de téléchargements est soumise à de nombreuses contraintes techniques avec lesquelles il faut conjuguer pour atteindre une précision et une qualité satisfaisantes. Nous présenterons d'abord les limites intrinsèques aux données et les biais introduits ou qui n'ont pas pu être éliminés lors du nettoyage des données, puis nous terminerons par les limites générales du projet.

1. Il aurait été intéressant d'inclure à notre étude les données de téléchargements depuis la création d'Érudit en 1998, mais les logs ne sont malheureusement pas disponibles. Nous n'avons, en outre, pas accès à six années complètes, ce dont il faut tenir compte dans nos résultats lorsque nous souhaitons montrer l'évolution de l'usage d'une année à l'autre.

2. Étant donné qu'il n'a pas été possible d'identifier la date de mise en ligne des numéros de revues, nous avons dû utiliser, pour les articles publiés depuis le 1er avril 2010, la date du premier téléchargement d'un article dans un numéro donné pour obtenir la date de son passage au libre accès et, pour les articles publiés avant le 1er avril 2010, la date de publication donnée par la revue. Pour contourner ce biais non négligeable dans notre analyse de l'effet du passage au libre accès sur l'usage des articles, nous avons dû nous restreindre aux articles mis en ligne entre 2011 et 2015, dont le calcul de l'embargo est fiable.

3. Pour des raisons de commodité, nous avons considéré que l'ensemble des articles d'un numéro possède le format de diffusion (format PDF seulement, ou encore format PDF et HTML) de la majorité. Nous perdons donc légèrement en précision dans le traitement des données, au moment de déterminer quels logs se terminant par « .html » exclure.

4. Le calcul de l'âge des articles téléchargés et le calcul de la fin de la période d'embargo est arrondie à l'année près, mais il aurait été beaucoup plus précis de connaître le jour et le mois de la publication des articles par les revues, de même que le jour et le mois de la mise en ligne des articles sur le site Web de la plateforme.

5. Il est possible qu'un nombre de téléchargements par des humains ait été éliminé lors de l'exclusion de robots, mais il s'agit sans doute d'un nombre négligeable. À l'inverse, il est également possible qu'un certain nombre de robots malicieux nous ait échappé. Nous pensons toutefois que le nombre de faux positifs et de faux négatifs n'influencera pas les tendances générales sur l'ensemble des données.

6. Même si les fuseaux horaires nous ont permis de calculer l'heure locale des téléchargements au moment de la géolocalisation, nous n'avons pas tenu compte des changements d'heure entre l'heure d'été et l'heure d'hiver pour les différents pays.

7. Les articles en traitement complet, dont il existe une version PDF et une version HTML du texte intégral, sont surreprésentés lorsqu'un utilisateur consulte la page Web d'un

article en traitement complet, puis télécharge la version PDF, car ces deux opérations comptent comme des téléchargements distincts. Les données contenues dans les fichiers de logs d'Érudit ne rendent pas possible l'identification de sessions de recherche, qui permettrait de regrouper les téléchargements d'un même usager dans un laps de temps donné, car les usagers n'ont pas à se connecter pour accéder aux articles de la plateforme.

8. Notre étude ne permet de nous prononcer que sur l'usage des articles de revues savantes sur la plateforme Érudit uniquement, et non pas l'usage de ces articles de manière globale. En effet, il est possible que les articles de la collection d'Érudit en libre accès aient été moissonnés par d'autres répertoires, ou encore que certains aient été auto-archivés sur un dépôt institutionnel ou sur le site personnel de chercheurs, mais nos données ne nous permettent pas de savoir si un article a été téléchargé depuis une autre source.

9. Les fichiers de logs d'Érudit n'enregistrent aucune donnée démographique sur les usagers qui aurait pu enrichir notre étude, comme l'ont proposé Nicholas et ses collaborateurs. Ces derniers ont mis au point une méthode, nommée « *deep log analysis* », pour enrichir leur analyse d'usage avec certaines données démographiques, telles que l'occupation ou le statut académique des usagers (Nicholas et al., 2005, p. 1446).

10. Ainsi qu'il a été abordé dans la revue de littérature, dans la section sur les données de téléchargements, un téléchargement ne mène pas nécessairement à une lecture. Il s'agit donc moins d'un indicateur du lectorat que d'un indicateur de l'intérêt suscité par un article.

11. Mais surtout, notre étude porte sur le contexte francophone des SSH au Québec, et plus précisément sur les collections d'Érudit. Il faut, par conséquent, demeurer prudent quant à la généralisabilité des résultats à d'autres contextes.

Maintenant que la méthodologie (en particulier, les décisions qui ont été prises quant au traitement des données) et que les principales limites du projet ont été exposées, le chapitre suivant présentera les résultats obtenus.

Chapitre 3 : Résultats

Ce chapitre est organisé en trois grandes parties. Les deux premières parties présentent les résultats liés à notre analyse exploratoire de l'usage de la plateforme Érudit, tandis que la troisième partie synthétise les résultats liés à l'impact du libre accès sur l'usage des articles de la plateforme auprès de la communauté universitaire.

Une grande variété de questions est abordée dans la première partie afin de dresser un portrait global de l'usage des articles de revues savantes sur la plateforme et d'explorer les possibilités offertes par l'indicateur webométrique des données de téléchargements. Cette analyse exploratoire permettra, en d'autres termes, de poursuivre notre réflexion sur cet indicateur – réflexion qui a été amorcée dans le chapitre précédent et qui faisait état des limites méthodologiques de cet indicateur, principalement en ce qui a trait au traitement des données. Dans la deuxième partie, qui vise à compléter le portrait global de l'usage des articles sur Érudit, nous nous concentrerons sur les habitudes de téléchargement des usagers selon différentes échelles de temps : par heure du jour, par jour de la semaine, par mois, par saison et par année. Nous comparerons les habitudes de téléchargement des usagers provenant des principaux pays qui utilisent la plateforme Érudit, ce qui nous permettra d'entrevoir certaines différences culturelles et de discuter des enjeux liés à la conciliation entre la vie personnelle et la vie professionnelle dans le monde académique. La troisième partie s'intéressera à l'impact du libre accès sur les téléchargements. Nous comparerons d'abord le nombre de téléchargements obtenus avant et après la fin de la période d'embargo pour des articles mis en ligne au même moment, en vue d'en dessiner la courbe d'usage dans le temps. Les différences entre les pays, les continents et les disciplines seront présentées, et les revues en accès différé seront confrontées à celles en libre accès complet. Suite à cela, nous observerons la proportion de téléchargements d'articles en libre accès, en tenant compte du nombre réel d'articles en libre accès dans la collection d'Érudit (qui ont donc le potentiel d'être téléchargés), bien plus nombreux que ceux en accès restreint.

Notons que la première partie des résultats est basée sur les données de téléchargements de l'ensemble des articles de la plateforme (sauf indication contraire, dans les cas où il était nécessaire de nous restreindre aux articles dont la date de mise en ligne sur le

site d'Érudit est connue). Les deux autres parties sont basées uniquement sur la période allant de 2011 à 2015 inclusivement.

Portrait global de l'usage des articles sur Érudit

Un article scientifique sur Érudit, en moyenne, est téléchargé 433,08 fois. Pour l'ensemble des logs colligés de 2010 à 2015, l'étendue va de 0 à 78 134 téléchargements pour un même article, la médiane est de 138 et le mode est de 0 ; en effet, 1 996 articles sur un total de 91 016 ne sont jamais téléchargés, et 236 sont téléchargés une fois seulement. Depuis 2011, le nombre de téléchargements sur la plateforme augmente légèrement, ainsi que l'illustre la courbe de tendance linéaire de la figure 4. Notons que l'importante baisse observée en 2014 s'explique par le fait que Google a cessé de référencer les articles de la collection pendant plusieurs mois, ce qui a considérablement affecté leur usage.

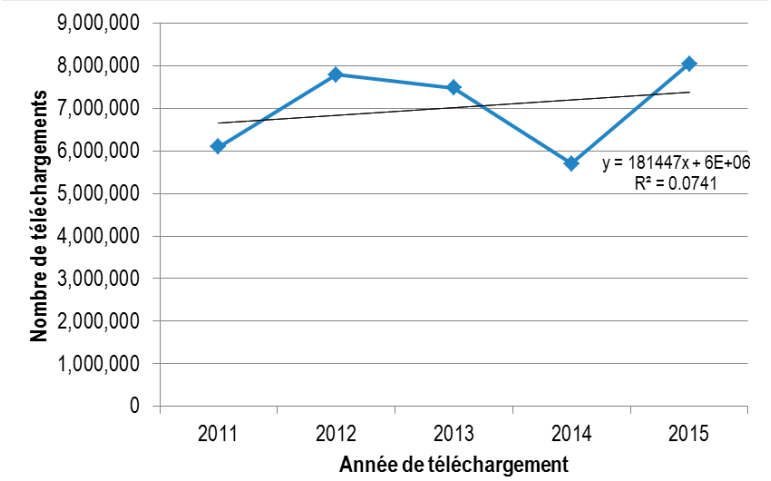

Figure 4. Nombre de téléchargements par année (n = 35 116 136[72])

Si le nombre de téléchargements annuel augmente, c'est également le cas pour le nombre d'articles contenu dans la collection. C'est pourquoi il est plus significatif d'examiner l'évolution dans le temps du nombre moyen de téléchargements par article, ce que nous ferons dans un instant.

[72] Le nombre de téléchargements dans cette figure est plus petit que le nombre total de téléchargements dans notre jeu de données (couvrant la période du 1er avril 2010 au 31 décembre 2015), car l'année 2010, incomplète, a été exclue.

Distribution des téléchargements par discipline

Après avoir dressé un portrait général des revues savantes contenues dans la plateforme Érudit et des disciplines couvertes, dans la section « Source des données » du chapitre précédent, nous présentons ici les résultats liés à notre analyse des données de téléchargements par grandes disciplines, disciplines et spécialités. En termes de nombre de téléchargements, les articles en sciences sociales et humaines (SSH) l'emportent haut la main avec 89,63% des téléchargements, contre 10,37% des téléchargements pour les sciences naturelles et le génie (SNG). Les dix spécialités les plus téléchargées sont, dans l'ordre : la littérature, les langues et la linguistique, la sociologie, l'éducation, le travail social, l'histoire, le management, la géographie, la psychiatrie, de même que l'anthropologie et l'archéologie, où seule la psychiatrie appartient aux SNG. Cela est, bien évidemment, dû à la couverture de la plateforme, dans laquelle 95,05% des articles de la collection appartiennent à la grande famille des SSH.

Si l'on considère plutôt le nombre moyen de téléchargements par article, on constate que les articles en SNG sont 2,21 fois plus téléchargés que ceux en SSH (nombre moyen de 173,75 téléchargements/article contre 78,66 pour les cinq ans de logs), exception faite d'une revue en sociologie qui a atteint la valeur maximale de 685,88 téléchargements par article en 2012. Toutes années confondues, les revues qui ont obtenu les plus importants nombres de téléchargements par article viennent des spécialités sociologie, psychiatrie, travail social et littérature. La figure 5 permet de démontrer que le nombre moyen de téléchargements par article reste stable d'une année à l'autre en SSH, mais qu'il tend à baisser en SNG.

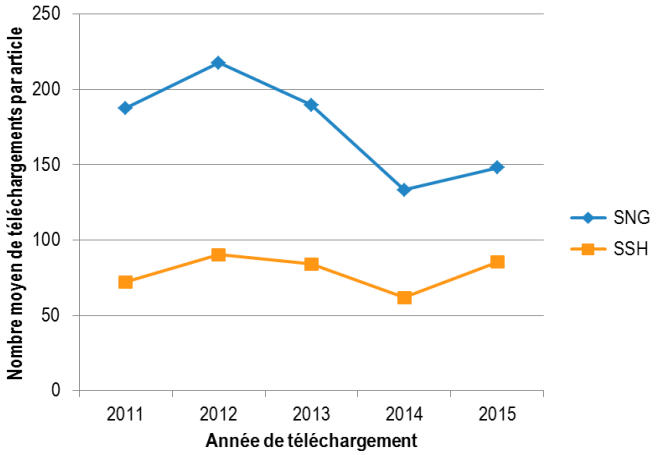

Figure 5. Nombre moyen de téléchargements par article pour les sciences sociales et humaines et pour les sciences naturelles et génie de 2011 à 2015 (n = 35 113 672[73])

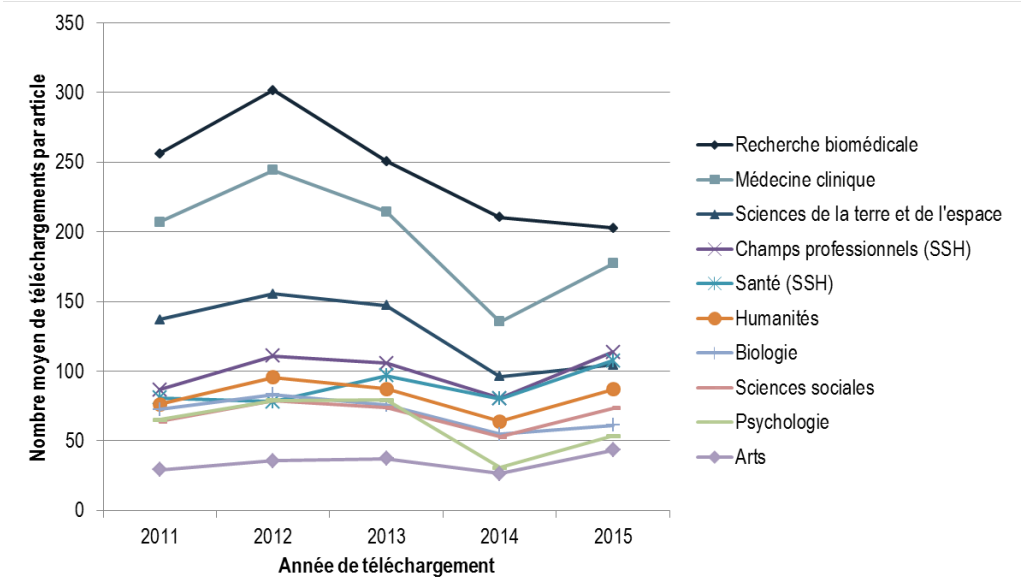

Figure 6. Nombre moyen de téléchargements par article par discipline de 2011 à 2015 (n = 35 113 672)

La figure 6 permet d'ordonner les disciplines par ordre croissant d'usage. Ainsi, les courbes au bas de la figure représentent les disciplines ayant le plus petit nombre moyen de téléchargements par article (p. ex. arts, psychologie, sciences sociales, biologie) et les courbes

[73] Le nombre de téléchargements comptés dans les figures 5 et 6 est légèrement plus petit que dans la figure 4, car dans quelques cas, les abréviations des titres de revues qui contiennent des caractères anormaux (généralement des caractères spéciaux qui ne devraient pas être là) ne permettent pas de les associer à une discipline.

en haut, les disciplines les plus utilisées (p. ex. recherche biomédicale et médecine clinique). Dans la légende, les quatre disciplines en partant du haut appartiennent à la famille des SNG, tandis que les autres appartiennent aux SSH.

La figure 7 montre la distribution des téléchargements par spécialité. Une distribution similaire a pu être observée pour les téléchargements par revues, mais la pente est moins abrupte, ce qui signifie que l'écart est plus petit entre les revues qui obtiennent le plus grand nombre de téléchargements et celles qui en obtiennent le moins, comparativement à l'écart entre les spécialités les plus téléchargées et les moins téléchargées.

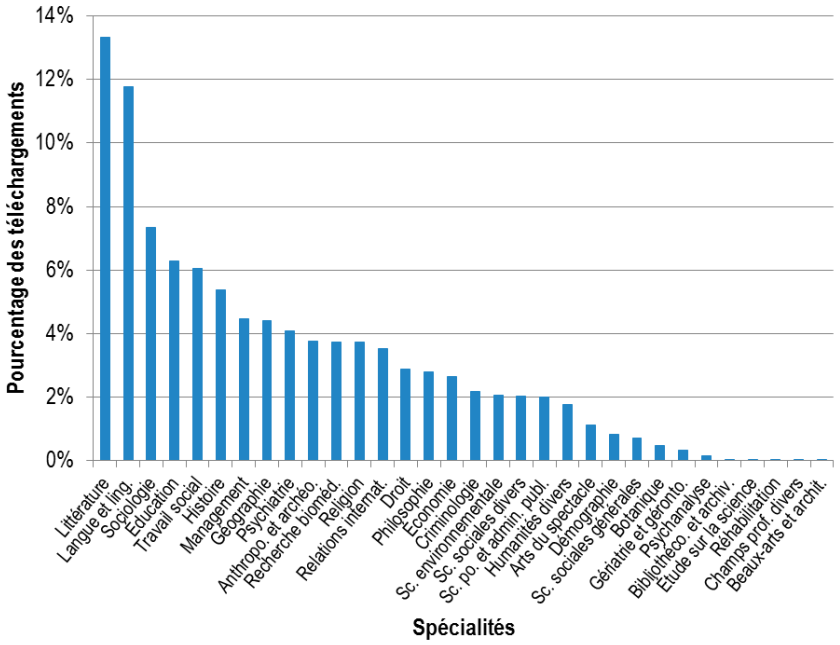

Figure 7. Distribution des téléchargements par spécialité (n = 39 434 833[74])

Âge des articles téléchargés

L'âge moyen des articles téléchargés sur la plateforme Érudit est de 15 ans, ce qui peut sembler assez élevé. Rappelons toutefois que les publications savantes en SSH ont une plus grande durée de vie que celles en sciences naturelles (p. ex. Houghton, 1975, p. 109-110; Line, 1993, p. 667; Nicholas et al., 2005, p. 1443). En outre, une étude a obtenu, pour les articles en

[74] Dans ce cas-ci, les téléchargements de l'année 2010 sont inclus, mais, comme c'était le cas dans les figures 5 et 6, les articles téléchargés dont la discipline n'a pas pu être identifiée ont été exclus.

SNG vers 2005, un âge médian de la littérature citée de 7 ans et un âge moyen de 11,5 ans (avec une fenêtre de citation de 100 ans) (Larivière, Gingras et Archambault, 2008, p. 290-291). Il ne nous paraît donc pas surprenant que la moyenne soit si élevée pour Érudit, d'autant plus que la barrière mobile limite les téléchargements d'articles très récents, comme nous le verrons plus loin. La médiane, quant à elle, est de 12 ans, et le mode est de 7 ans. Dans la figure 8, on voit que la moyenne est restée assez stable depuis 2010, augmentant à 16 ans en 2012 et 2013, diminuant à 13 ans en 2014, puis remontant à 16 ans en 2015.

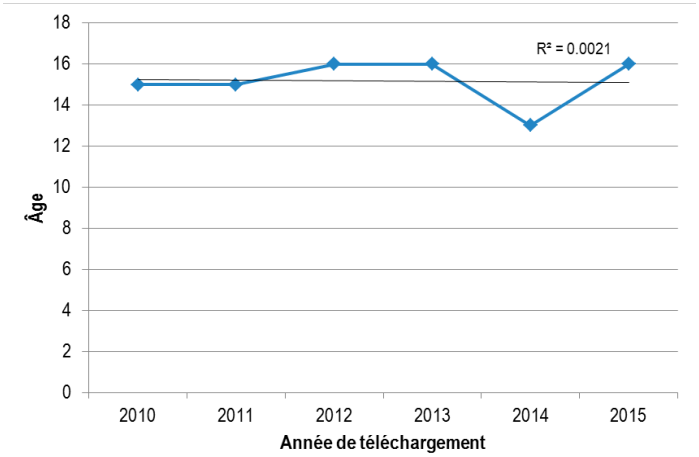

Figure 8. Âge moyen des articles téléchargés de 2010 à 2015 (n = 39 437 659[75])

La figure 9 montre que la distribution des âges[76] présente une courbe asymétrique à droite pour l'ensemble des téléchargements (rappelons que la moyenne étant une mesure de tendance centrale, elle est bien moins représentative d'une distribution asymétrique que d'une distribution normale). Si l'on compare les grandes disciplines, l'âge moyen des articles téléchargés en SSH est de 16 ans, contre 11 ans pour les articles en SNG, mais les distributions sont assez différentes dans les deux cas. Les publications en SSH sont âgées de 0 à 93 ans, tandis que celles en SNG sont âgées de 0 à 39 ans seulement, ce qui s'explique par l'obsolescence des publications dans ces grandes disciplines. En outre, pour les SSH, le mode de la distribution est de 5 ans, soit l'âge d'à peine plus de 5% des publications téléchargées. Étonnamment, le mode des SNG est atteint à 8 ans (l'âge de plus de 10% des publications),

[75] Tous les téléchargements du 1er avril 2010 au 31 décembre 2015 ont été comptés.
[76] La collection d'Érudit ne contient pas d'articles publiés avant 1922. Les données de logs contenant certaines erreurs, les années de publication antérieures à 1922 ont été exclues manuellement.

mais cela s'explique par le fait que la seconde revue la plus téléchargée dans ces disciplines a cessé d'être publiée en 2006.

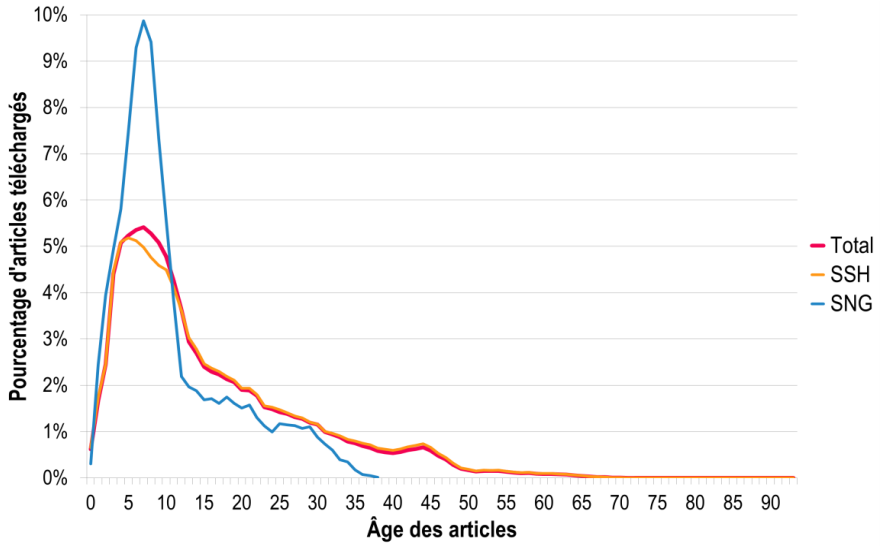

Figure 9. Âge des articles téléchargés sur Érudit au total, pour les sciences sociales et humaines, ainsi que pour les sciences naturelles et génie (n = 39 436 791[77])

Provenance des usagers

La géolocalisation est complétée avec succès dans 99,18% des cas à l'échelle des continents, dans 98,64% des cas pour les pays, dans 92,70% des cas pour les fuseaux horaires, dans 68,12% des cas pour les régions et dans 68,09% des cas seulement pour les villes. En outre, ces dernières sont assez peu précises ; on doit donc comprendre « Montréal » comme Montréal et ses environs. Les résultats présentés excluent les téléchargements où le continent, le pays, le fuseau horaire, la région ou la ville (selon le cas) ne sont pas déterminés.

Comme le montre le tableau 2, les usagers en provenance du Canada sont, sans surprise, ceux qui effectuent le plus grand nombre de téléchargements sur Érudit (29,42% du nombre total de téléchargements). On aurait toutefois pu s'attendre à ce qu'une plus grande proportion de téléchargements provienne du Canada, car la France suit ce pays d'assez près (22,63% des téléchargements). En troisième position viennent les États-Unis, qui sont déjà

[77] Les téléchargements qui ont eu lieu en 2010 ont été conservés. En outre, les articles téléchargés dont la discipline n'a pas pu être identifiée ont été comptés dans la courbe « Total », ce qui explique que le nombre de valeurs de cette figure soit légèrement plus grand que celui de la figure 7.

loin derrière le Canada et la France (6,43% des téléchargements). Viennent ensuite, dans l'ordre, la Chine, l'Algérie, le Maroc, l'Allemagne, la Belgique, la Tunisie, ainsi que le Royaume-Uni. On voit la prépondérance des pays d'Europe de l'Ouest (France, Royaume-Uni, Allemagne, Belgique), de même que des pays où le français est une langue officielle (Canada, France, Belgique) ou encore où il est couramment parlé, c'est-à-dire les pays d'Afrique du Nord (Algérie, Maroc, Tunisie). La position occupée par les États-Unis et la Chine se justifie certainement par la grande proportion de chercheurs qui sont issus de ces deux pays sur la scène mondiale. Par ailleurs, les usagers en provenance de Montréal et de Québec effectuent 16,96% des téléchargements. Étrangement, Shenzhen occupe la troisième place, assez près de Québec.

Tableau 2. Dix pays qui téléchargent le plus (n = 38 901 921) et dix villes qui téléchargent le plus (n = 26 852 807[78])

	Pays	Ratio
1	Canada	29.42%
2	France	22.63%
3	États-Unis	6.43%
4	Chine	4.72%
5	Algérie	4.32%
6	Maroc	3.47%
7	Allemagne	3.27%
8	Belgique	2.71%
9	Tunisie	1.81%
10	Royaume-Uni	1.56%

	Villes	Ratio
1	Montréal, Canada	11.92%
2	Québec, Canada	5.03%
3	Shenzhen, Chine	3.75%
4	Paris, France	3.45%
5	Ottawa, Canada	2.33%
6	Gatineau, Canada	1.65%
7	Sherbrooke, Canada	1.48%
8	Toronto, Canada	1.01%
9	Laval, Canada	1.01%
10	Chicago, États-Unis	0.89%

Au Canada, enfin, le tableau 3 confirme que 80,50% des téléchargements proviennent du Québec et 13,84% de l'Ontario. Viennent ensuite, dans l'ordre, le Nouveau-Brunswick, la Colombie Britannique, l'Alberta, le Manitoba, la Nouvelle-Écosse, la Saskatchewan, Terre-Neuve-et-Labrador, l'Île-du-Prince-Édouard, les Territoires du Nord-Ouest, le Yukon et le Nunavut. La distribution des téléchargements par pays, par villes et par provinces et territoires du Canada est, dans les trois cas, non paramétrique et présente une asymétrie à droite.

[78] Le pourcentage de téléchargements en provenance des dix principaux pays se base sur le nombre total de téléchargements effectués entre le 1er avril 2010 et le 31 décembre 2015, à l'exception de ceux dont le pays n'a pas pu être géolocalisé. Il en va de même pour les villes.

Tableau 3. Le pourcentage d'articles téléchargés par provinces et territoires du Canada (n = 10 795 008[79])

Provinces et territoires du Canada	Ratio
Québec	80.50%
Ontario	13.84%
Nouveau Brunswick	1.79%
Colombie Britannique	1.13%
Alberta	1.03%
Manitoba	0.59%
Nouvelle Écosse	0.57%
Saskatchewan	0.30%
Terre-Neuve-et-Labrador	0.15%
Île-du-Prince-Édouard	0.06%
Territoires du Nord-Ouest	0.01%
Yukon	0.01%
Nunavut	0.01%

Référents les plus usités

Grâce au référent, il est possible de savoir à partir de quel site Web un usager a téléchargé un article de la collection d'Érudit (p. ex. le site de la plateforme, un moteur de recherche, une boîte de courriels, une page Web d'institution universitaire, un site de médias sociaux, etc.). L'absence de référent peut signifier que l'usager a téléchargé un article à partir d'une fenêtre de navigation privée, que l'article avait été enregistré au préalable dans les favoris de l'usager, ou encore qu'il s'agit d'un robot. Notre technique de détection de robots, expliquée dans le chapitre précédent, tient compte de cette particularité.

Les quelque 39 millions de téléchargements effectués depuis 2010 proviennent de 34 461 référents différents. Fait intéressant à noter, le moteur de recherche de Google (tous domaines confondus, c'est-à-dire : www.google.com, www.google.ca, www.google.be, www.google.it, www.google.fr, etc.) est 1,62 fois plus utilisé que la plateforme Érudit (15 794 236 téléchargements obtenus contre 9 740 036 pour Érudit). De surcroît, si l'on additionne l'ensemble des services de Google, soit le moteur de recherche généraliste, Google Scholar, Google Images et Google Translate, on obtient que le géant du Web est 1,82 fois plus utilisé qu'Érudit (avec 17 753 479 téléchargements). La figure 10 montre que, si l'on exclut les référents vides (c'est le cas pour 24,05% des téléchargements), les moteurs de recherche

[79] Le pourcentage de téléchargements en provenance de chaque province et territoire se base sur le nombre total de téléchargements effectués au Canada du 1er avril 2010 au 31 décembre 2015.

les plus connus viennent en tête, de même que le site de la plateforme Érudit et celui de la base de données Repère.

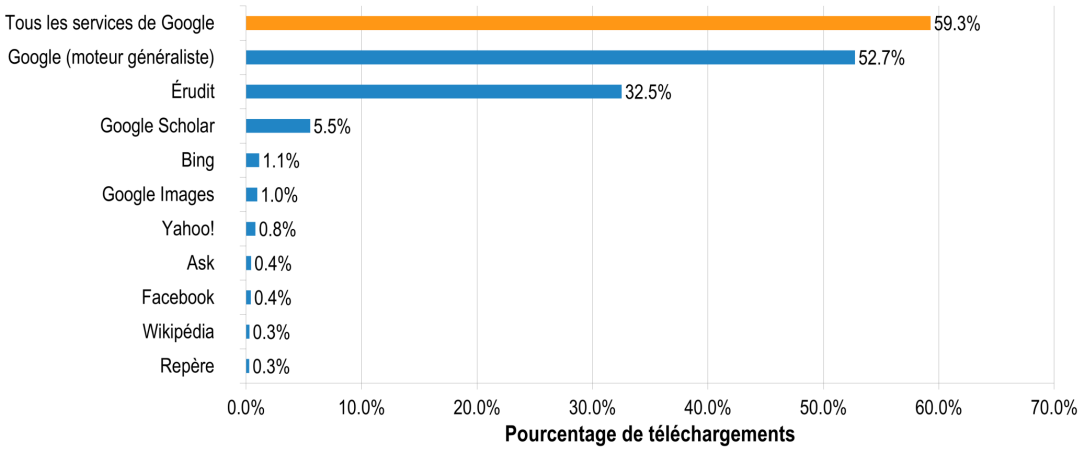

Figure 10. Les dix référents les plus courants, ainsi que la somme des téléchargements effectués depuis l'un des services de Google (en orange) (n = 29 951 265[80])

Les partages sur le réseau social Facebook, les citations sur l'encyclopédie libre Wikipédia, la plateforme Conduit (dont les barres d'outils sont considérées comme malveillantes puisqu'elles sont souvent installées à l'insu des utilisateurs et modifient certains paramètres du navigateur tels que le moteur de recherche par défaut), ainsi que le site de l'Université de Montréal entraînent aussi un certain nombre de téléchargements.

[80] Cette figure inclut tous les téléchargements effectués entre le 1er avril 2010 et le 31 décembre 2015 dont le référent a pu être identifié.

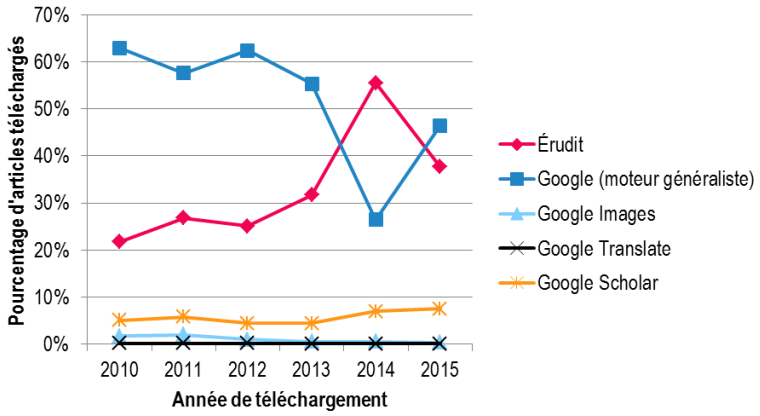

Figure 11. Pourcentage d'articles téléchargés depuis la plateforme Érudit et depuis l'un des services de Google de 2010 à 2015 (n = 27 493 515[81])

La figure 11 permet de comparer le site d'Érudit aux différents moteurs de recherche de Google dans le temps. Après une forte chute de 2010 à 2011, la place occupée par le site d'Érudit a augmenté de 2011 à 2014, mais a connu une baisse en 2015. Le moteur de recherche généraliste de Google, quant à lui, a connu une baisse de 2010 à 2014, puis une hausse en 2015. Rappelons que Google a cessé de référencer les articles de la collection en 2014. Les autres services de Google, quant à eux, sont restés stables depuis 2010. Pour ce qui est des articles en SSH et en SNG, la proportion de téléchargements en provenance de Google est comparable (environ 53% dans les deux cas), mais les téléchargements en provenance du site d'Érudit sont bien moins nombreux dans le cas des SNG (25,10% des téléchargements en SNG contre 33,43% pour les SSH). Cela signifie que la distribution des référents est plus concentrée dans les SSH.

Appareils utilisés par les usagers

Les ordinateurs sont utilisés dans 93,69% des téléchargements, contre 4,33% pour les téléphones portables et 1,98% pour les tablettes (en excluant les 1,36% des cas où les logs ne permettent pas de reconnaître l'appareil utilisé par l'usager). Depuis 2010, la proportion de téléphones a augmenté d'environ 9% et celle de tablettes de 4%, au détriment des ordinateurs qui ont baissé de 12%. Parmi les usagers qui ont téléchargé des publications sur Érudit depuis

[81] Cette figure inclut tous les téléchargements effectués entre le 1er avril 2010 et le 31 décembre 2015 dont le référent est le site Web d'Érudit ou l'un des services de Google.

un ordinateur, 84,15% ont utilisé l'un des systèmes d'exploitation de Windows, 13,65% travaillent sous Mac et 2,20% ont utilisé un autre système d'exploitation. Enfin, 30,77% des usagers utilisent le navigateur Internet Explorer, 27,63% Firefox, 23,25% Chrome et seulement 7,79% Safari, la faible proportion pour ce dernier navigateur étant liée au petit nombre d'usagers qui utilisent Mac.

Les résultats présentés dans la première partie sont d'un grand intérêt stratégique pour les membres de la direction d'Érudit. En effet, il leur est essentiel d'en apprendre davantage sur les usagers pour améliorer leurs services et optimiser la plateforme. Par exemple, la provenance géographique des usagers pourrait permettre d'orienter l'ajout de nouveaux corpus en des langues autres que le français et l'anglais; les appareils, navigateurs et systèmes d'exploitation utilisés indiquent quelles versions du site et quelles applications devraient être développées; le nombre moyen de téléchargements par article par discipline et par revue permet de déterminer quels contenus mettre de l'avant; l'importance respective des référents permet d'améliorer la visibilité et l'indexation des contenus; et ainsi de suite.

Habitudes de téléchargement des usagers

Dans cette partie, nous présentons les résultats liés aux habitudes de téléchargement des usagers sur la plateforme Érudit pour les trois pays qui effectuent 58,48% de tous les téléchargements, soit le Canada (29,42% des téléchargements, ou 11 443 710 téléchargements), la France (22,63%, ou 8 803 752 téléchargements) et les États-Unis (6,43%, ou 2 502 400 téléchargements)[82]. Différentes échelles de temps sont observées : l'activité quotidienne et hebdomadaire en ligne, de même que l'activité au fil de l'année académique, c'est-à-dire par mois et par saison. Dans un dernier temps, une perspective diachronique est adoptée pour comparer les années académiques entre elles.

[82] Même si la proportion de téléchargements en provenance des États-Unis est bien moindre que celle observée au Canada et en France, nous avons les avons tout de même inclus dans cette partie de l'analyse à cause de leur intérêt grandissant envers la collection d'Érudit (Henry, 2016). Cet intérêt a été constaté par le renouvellement des abonnements de plusieurs bibliothèques prestigieuses (Harvard Library, Yale University Library, Duke University Libraries, University of Washington Libraries, University of Vermont Libraries et New York Public Library) aux articles en accès restreint de la plateforme depuis trois années consécutives (Henry, 2016).

Dans cette partie n'ont été conservés que les téléchargements qui ont pu être géolocalisés (92,70% des cas à l'échelle des fuseaux horaires) et, par conséquent, dont on connaît l'heure locale. L'année 2010 a également été exclue, à cause des quatre mois de fichiers de logs manquants (janvier, février, mars et août) ; seulement les cinq années complètes de 2011 à 2015 sont observées. Pour plus de précision, l'ensemble des requêtes effectuées dans cette partie – y compris le calcul de l'année de téléchargement – sont basés sur l'heure locale (dans les deux autres parties de nos résultats, nous avons calculé l'année de téléchargement à partir de l'heure de Montréal pour couvrir l'ensemble des données). En outre, dans le cas des figures 13, 14 et 15[83], pour les quelques jours manquants dans les fichiers de logs (les 15, 17 et 18 décembre 2011, les 25, 26 et 27 août 2012, ainsi que les 16, 17 et 24 février 2013), de même que pour les jours incomplets parce qu'affectés par les fichiers manquants à cause de l'heure locale (le jour immédiatement avant et celui immédiatement après chaque fichier manquant), un nombre de téléchargements moyen a été calculé à partir des téléchargements de quatre jours, en prenant les deux mêmes jours de la semaine avant le fichier affecté, et les deux mêmes jours de la semaine après. Ainsi, pour un mercredi affecté, la moyenne a été calculée à partir des deux mercredis précédents et des deux mercredis suivants.

Cette partie de notre analyse présuppose que les téléchargements reflètent une certaine forme de dévouement au travail, mais les téléchargements sous-estiment certainement la quantité réelle de travail effectué par les chercheurs les soirs, les week-ends et les jours fériés, puisqu'un téléchargement à un temps donné peut mener à une lecture ultérieure (Cabanac et Hartley, 2013, p. 2).

[83] À l'échelle des heures de la journée, soit pour la figure 12, il n'a pas été jugé nécessaire de compenser pour les jours manquants. C'est plutôt à l'échelle des journées, des mois et des saisons que cela peut introduire un biais, car cela peut mener à la sous-représentation des téléchargements effectués certains jours de la semaine et certains mois où davantage de fichiers sont manquants. C'est pourquoi le nombre de téléchargements inclus dans la figure 12 est légèrement plus petit que celui des figures 13, 14 et 15.

Activité quotidienne

La figure 12 montre que les usagers qui proviennent du Canada[84] commencent leur journée de travail légèrement plus tôt que les Français (début de la hausse vers 5h pour le Canada, contre 6h pour la France), mais ces derniers continuent à travailler plus tard en fin d'après-midi et en soirée (accélération du déclin à compter de 21h pour le Canada, contre 22h pour la France). La courbe canadienne présente deux pics, le premier à 11h (heure du jour qui obtient 8,24% des téléchargements) et le second à 14h (qui obtient 8,55% des téléchargements). La courbe française, quant à elle, comporte aussi deux pics : à 11h (qui obtient 7,13% des téléchargements) et entre 15h et 16h (respectivement 8,02% et 7,98% des téléchargements). Le sommet atteint en après-midi par la France est beaucoup plus élevé que celui en avant-midi, à la différence du Canada où les deux pics ont pratiquement la même importance. La valeur minimale est atteinte à 5h pour le Canada (0,32% des téléchargements) et pour la France (0,56% des téléchargements). Par ailleurs, les Français sont plus actifs que les Canadiens pendant la nuit. Pour ces deux pays, on constate également une diminution non négligeable du nombre de téléchargements sur l'heure du dîner[85] (à 12h pour le Canada et entre 12h et 13h pour la France) ; le déclin est toutefois plus prononcé et d'une durée plus longue en France. Il en va de même pour le souper[86], où les téléchargements diminuent pour les trois pays, les Canadiens étant ceux qui prennent leur repas le plus tôt (vers 18h), suivis par les Américains (vers 19h) et par les Français (vers 20h).

[84] Nous avons ajouté le Québec pour montrer que cette province suit la même tendance que l'ensemble du Canada. Pour cette raison, nous n'avons pas jugé nécessaire d'ajouter une courbe pour le Québec sur les autres figures portant sur les habitudes de téléchargement des usagers.
[85] Au sens de « repas du midi ».
[86] Au sens de « repas du soir ».

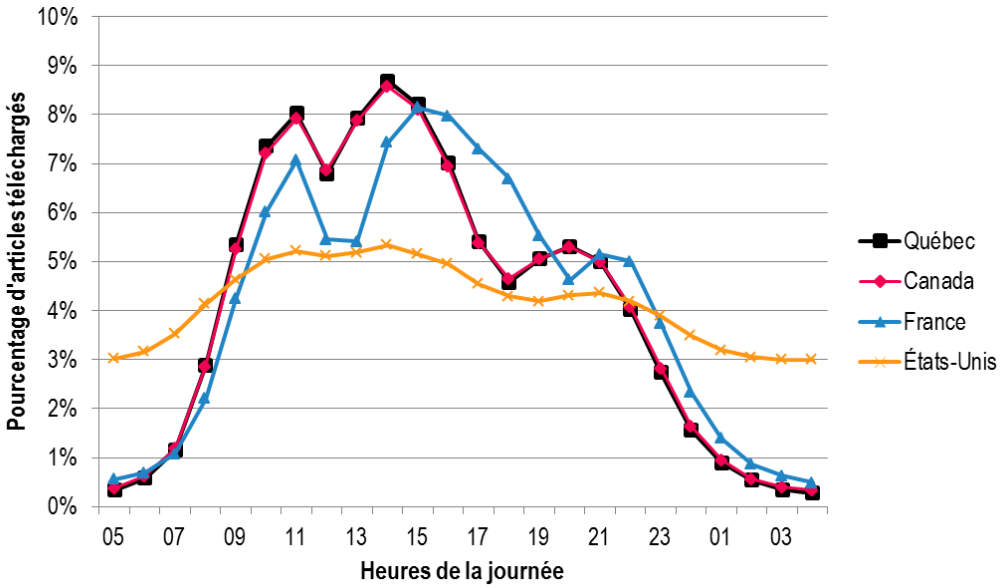

Figure 12. Proportion de téléchargements par heure du jour pour le Canada, la France et les États-Unis (n = 19 318 374)

Chose étonnante, confirmée par l'étude de Wang et ses collaborateurs (Wang et al., 2012), les habitudes de téléchargement des usagers aux États-Unis suggèrent un emploi du temps très différent de celui des Canadiens et des Français. Wang et ses collaborateurs ont montré que les Américains sont très actifs la nuit, y compris le week-end, ce que confirment nos résultats. Les téléchargements sont, en effet, assez constants au fil de la journée et de la nuit. Si l'on compare la valeur maximale (5,45% des téléchargements quotidiens obtenus à 15h) et la valeur minimale (2,87% des téléchargements à 3h), on remarque à quel point la différence entre le nombre de téléchargements le jour et la nuit est faible pour ce pays. De surcroît, l'heure du dîner ne semble qu'avoir peu d'effet sur l'activité des usagers américains ; on remarque un léger ralentissement à 10h, puis à 13h, ce qui laisse penser que les Américains prennent leur pause à des moments différents, ou encore qu'ils continuent de travailler tout en mangeant. Par contre, l'effet du souper est bien visible, avec un déclin net à 19h, suivi d'une rapide croissance à 20h, puis d'un très lent déclin en fin de soirée et pendant la nuit.

Activité hebdomadaire

À l'échelle des jours de la semaine, notre analyse des habitudes de téléchargement des usagers sur Érudit est en adéquation avec les résultats obtenus dans d'autres études. Encore

une fois, ainsi que l'illustre la figure 13, le Canada et la France présentent des tendances similaires, avec des usagers qui respectent une semaine de travail conventionnelle : la majorité des téléchargements sont effectués les jours de semaine. La journée la plus active chez les Canadiens est le lundi, qui obtient 22,56% des téléchargements hebdomadaires, suivi par le mercredi et le jeudi (respectivement 16,99% et 17,13% des téléchargements). La proportion de téléchargements diminue ensuite drastiquement du vendredi au dimanche, ce dernier jour obtenant seulement 8,93% des téléchargements. La France suit une courbe similaire, mais sans qu'il y ait une journée qui se démarque clairement par son nombre élevé de téléchargements. Dans l'ordre, le jeudi, le lundi et le mercredi se suivent de près (respectivement 17,62%, 16,94% et 16,27% des téléchargements hebdomadaires). En un mot, pour ces deux pays, le nombre de téléchargements est élevé le lundi, diminue légèrement le mardi, augmente de nouveau jusqu'au jeudi et, enfin, atteint son niveau le plus bas le week-end. Toutefois, contrairement au Canada où le nombre de téléchargements diminue de façon constante du vendredi au dimanche (on passe d'une proportion de 10,85% de téléchargements à 9,06%, puis à 8,93%), la France connaît une hausse d'activité le dimanche (11,71% des téléchargements), où le nombre de téléchargements est plus élevé que le samedi (9,78% des téléchargements), mais tout de même moins que le vendredi (14,05% des téléchargements).

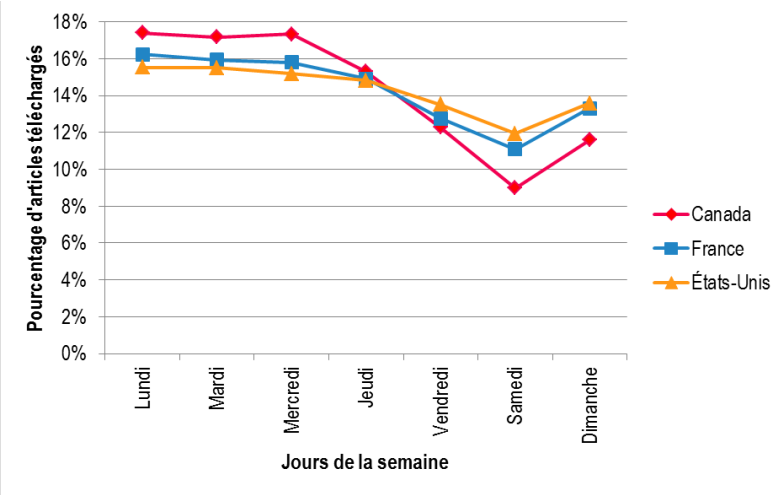

Figure 13. Proportion de téléchargements par jour de la semaine pour le Canada, la France et les États-Unis (n = 19 413 395)

Encore une fois, les États-Unis se démarquent par une plus grande homogénéité dans le nombre de téléchargements au fil des jours de la semaine (et des week-ends). En effet, l'écart

entre la valeur maximale (15,88% des téléchargements hebdomadaires sont effectués le jeudi) et la valeur minimale (12,11% des téléchargements le samedi) est assez faible. Les Américains sont même plus actifs le dimanche que le vendredi (où respectivement 14,45% et 13,76% des téléchargements sont effectués). Dans le cas des trois pays, la hausse des téléchargements qui se produit le dimanche s'explique sans doute par le phénomène du « *academic Sunday night* », où les chercheurs commencent à préparer leur semaine de travail.

Activité mensuelle et saisonnière

Les usagers au Canada sont particulièrement actifs de septembre à novembre et de février à mars, donc surtout en période de mi-session, ce qui pourrait s'expliquer par un usage intensif de la plateforme Érudit par des étudiants de premier cycle. Plus précisément, comme le montre la figure 14, la courbe canadienne augmente de septembre jusqu'à atteindre sa valeur maximale en novembre (qui obtient 14,27% des téléchargements mensuels), elle chute dramatiquement en décembre, avant de remonter jusqu'à atteindre un second pic en février (13,41% des téléchargements) et, enfin, de redescendre progressivement jusqu'en été, où la valeur minimale est atteinte en juillet (3,15% des téléchargements). En ce qui concerne la France, le nombre de téléchargements atteint un premier sommet en octobre (11,92% des téléchargements), décline jusqu'en décembre, remonte en janvier et atteint sa valeur maximale en mars (12,50% des téléchargements), pour ensuite chuter en avril, remonter très légèrement en mai et juin, puis descendre progressivement jusqu'à atteindre sa valeur minimale en août (3,60% des téléchargements). La tendance américaine est, encore une fois, plus stable, mais on peut tout de même percevoir des variations mensuelles. Ainsi, la valeur maximale est atteinte en octobre (10,48% des téléchargements), puis le nombre de téléchargements redescend doucement jusqu'en décembre, remonte pour atteindre un second sommet en février (9,76% des téléchargements), puis redescend pour atteindre sa valeur minimale en juillet (5,59% des téléchargements). Le déclin que l'on observe en décembre et janvier dans les trois pays est certainement causé par la période des Fêtes. En outre, le ralentissement qui commence en mars pour le Canada et les États-Unis, et en avril pour la France, est sans doute influencé par des événements spéciaux tels que des conférences, les vacances de Pâques, ou encore la semaine de relâche.

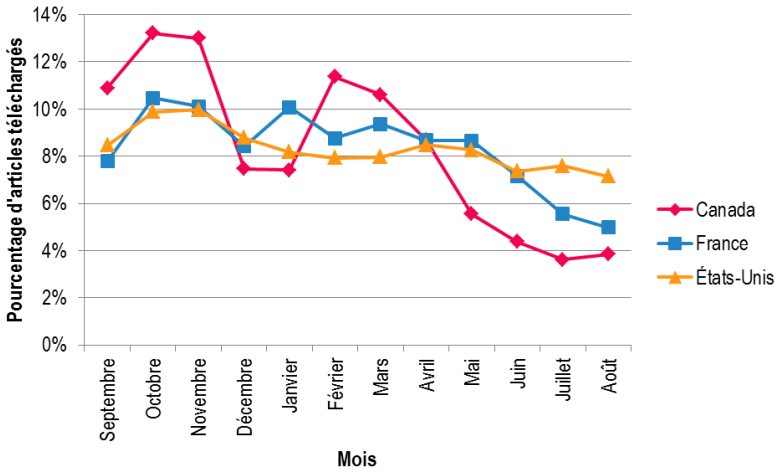

Figure 14. Proportion de téléchargements par mois pour le Canada, la France et les États-Unis (n = 19 413 395)

Si l'on s'intéresse maintenant au passage des saisons, le nombre de téléchargements pour les trois pays étudiés diminue graduellement au fil de l'année académique. La figure 15 montre que l'automne remporte le plus grand nombre de téléchargements et l'été le plus faible. Le Canada est le pays dont l'écart type est le plus important, avec 40,16% des téléchargements effectués en automne et seulement 14,90% en été. Vient ensuite la France, avec 32,29% des téléchargements en automne et 16,51% en été, puis les États-Unis, avec 29,66% des téléchargements en automne et 22,44% des téléchargements en été. Les usagers en provenance des trois pays sont, par conséquent, les plus actifs en automne sur la plateforme Érudit, et les moins actifs en été. En France, les valeurs sont très proches en automne et en hiver ; aux États-Unis, ce sont plutôt l'hiver, le printemps et l'été qui ont les valeurs les plus proches les unes des autres. En automne, ce sont les Canadiens qui téléchargent le plus ; en hiver, ce sont les Français ; au printemps et, surtout, en été, ce sont les Américains. Par ailleurs, durant la saison estivale, le Canada et la France téléchargent moins le soir, mais ce n'est pas le cas pour les États-Unis.

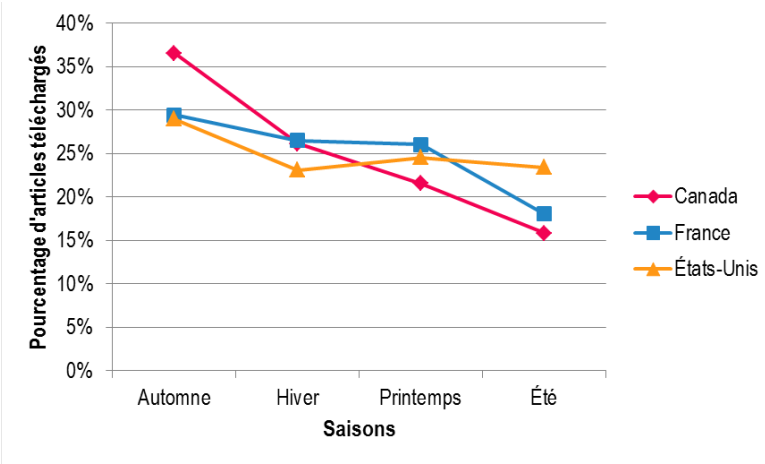

Figure 15. Proportion de téléchargements par saison pour le Canada, la France et les États-Unis (n = 19 413 395)

En comparant les habitudes de téléchargement des usagers d'une année à l'autre, nous avons constaté très peu de changement dans les tendances pour les trois pays. Seul détail à noter, les téléchargements par heure du jour sont légèrement plus lisses en 2015 pour le Canada. Cela pourrait s'expliquer par la durée de la période étudiée – cinq ans –, sans doute trop courte pour révéler une évolution dans les mœurs. À titre de comparaison, Cabanac et Hartley (2013) avaient pu observer une augmentation dans l'activité en ligne des chercheurs en lien avec le travail, mais sur une période allant de 2001 à 2012. Par ailleurs, si l'on compare les grandes disciplines, il n'y a pas de différence significative entre les habitudes de téléchargement entre les SSH et les SNG. Par contre, à l'échelle des disciplines en SSH, les téléchargements d'articles en santé semblent témoigner de plus saines habitudes de vie (moins de téléchargement la nuit, le week-end et l'été) que les autres disciplines, notamment les arts et les humanités.

Pour terminer, l'un des constats les plus surprenants de cette seconde partie de nos résultats est la grande constance dans les téléchargements en provenance des États-Unis, peu importe l'échelle de temps observée. Cette particularité pourrait peut-être s'expliquer par les types d'usagers que l'on rencontre en des proportions différentes selon les pays. Tandis que la plupart des téléchargements au Canada et, plus spécifiquement, au Québec sont effectués par des étudiants de premier cycle (davantage actifs durant l'année scolaire, mais moins pendant la période des Fêtes et en été), les téléchargements en France et sans doute plus encore aux États-

Unis sont peut-être effectués par des chercheurs, ce qui pourrait expliquer la proportion plus élevée de téléchargements en été aux États-Unis. Pour vérifier cette hypothèse, d'autres études devraient être menées sur les pratiques de travail des chercheurs selon le pays dans lequel ils résident, de même que sur l'équilibre entre la vie professionnelle et la vie personnelle dans le monde universitaire. Il est toutefois pertinent pour l'équipe d'Érudit de connaître les habitudes de téléchargement des usagers par provenance géographique pour déterminer, par exemple, les moments les plus opportuns pour procéder à la maintenance du site.

Impact du libre accès

Dans cette dernière partie de nos résultats, nous avons cherché à mesurer l'impact du libre accès sur les téléchargements. Au moment de la collecte des données, environ 95% des articles de la collection d'Érudit étaient disponibles en libre accès (Érudit, s.d.-b) et environ 98% des téléchargements correspondaient à des articles en libre accès. Cependant, n'ayant pas pu retracer rétrospectivement les dates de début et de fin de la barrière mobile des numéros de revue, rappelons que nous avons dû calculer cette date en utilisant l'année du premier téléchargement d'un article dans un numéro donné. Cela signifie que nos résultats ne concernent que les articles mis en ligne entre 2011 et 2015 pour lesquels il existe des données de logs, ce qui ne représente qu'une partie de la collection, d'une part, et ce qui sous-estime le nombre de téléchargements d'articles en libre accès, d'autre part. Le nombre de téléchargements sur lequel se base chaque figure a été indiqué.

Ainsi qu'il a déjà été précisé, il existe plusieurs manières de mesurer l'obsolescence des articles scientifiques en se basant sur le nombre de citations dans le temps (Larivière, Gingras et Archambault, 2008, p. 289). De prime abord, nous nous sommes inspirés des études diachroniques en fixant l'année de mise en ligne des articles, puis en observant les téléchargements subséquents de cet échantillon d'articles. Cela nous a permis d'examiner la courbe d'usage d'articles de revues en accès différé mis en ligne au même moment, avant et après la fin de l'embargo, puis de la comparer à la courbe d'usage d'articles de revue en libre accès complet mis en ligne au même moment. Dans un second temps, nous avons pris exemple sur les études synchroniques en fixant l'année de téléchargement à 2015 (afin d'inclure le plus grand nombre de téléchargements possible), mais en incluant l'ensemble des

articles mis en ligne entre 2011 et 2015. Ceci a pour but pour de confronter la proportion de téléchargements d'articles en libre accès à la proportion réelle d'articles en libre accès dans la collection d'Érudit (qui ont donc le potentiel d'être téléchargés).

Le libre accès en diachronie

La figure 16 montre l'évolution du nombre moyen de téléchargements par article pour les revues en accès différé, de même que pour les revues en accès complet. Pour les deux types de revues, de gauche à droite, la courbe d'usage est présentée pour les articles ayant été mis en ligne en 2011 et dont l'embargo se termine en 2013, pour les articles mis en ligne en 2012 dont l'embargo se termine en 2014, ainsi que pour les articles mis en ligne en 2013 dont l'embargo se termine en 2015. Dans les trois cas, les revues en libre accès complet ont un net avantage sur les autres : non seulement elles ont un nombre moyen de téléchargements par article bien plus élevé que les autres, mais encore elles atteignent leur pic d'utilisation dès la fin de la première année après leur mise en ligne, ce qui est loin d'être le cas pour les revues en accès différé. Les articles de ces dernières sont très peu téléchargés pendant la période d'embargo, puis leur usage se met à augmenter rapidement. Ils n'arrivent à rattraper le retard qu'ils ont acquis en début de vie qu'après quatre ans.

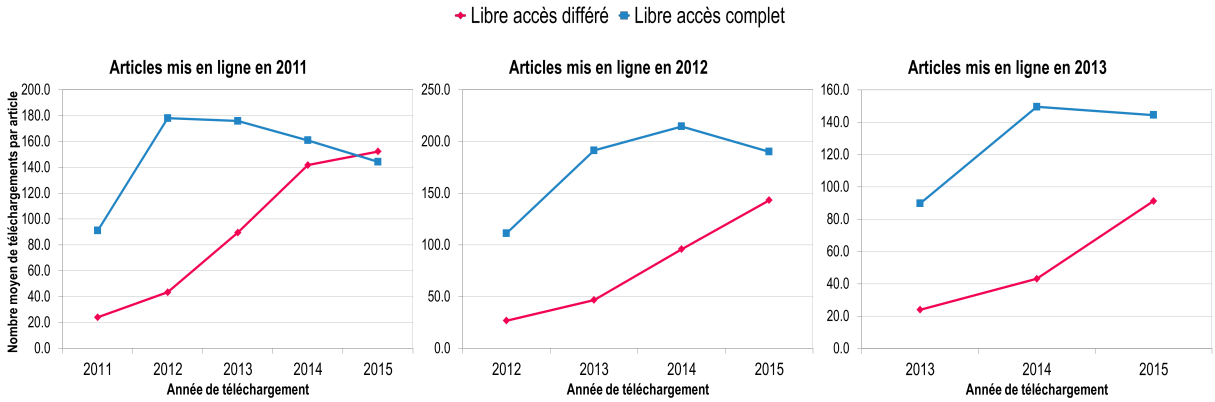

Figure 16. Nombre moyen de téléchargements par article pour les revues en libre accès différé et pour les revues en libre accès complet. De gauche à droite : articles mis en ligne en 2011 (n = 1 124 508) ; articles mis en ligne en 2012 (n = 843 113) ; articles mis en ligne en 2013 (n = 535 137)

Dans le tableau 4, les statistiques d'usage de l'échantillon des articles mis en ligne en 2011 prouvent que, de 2011 à 2015, le nombre moyen de téléchargements augmente pour les

deux types de revues, mais que celles en accès différé sont loin derrière les autres. Dans tous les cas, le nombre médian de téléchargements, plus petit que le nombre moyen de téléchargements, montre que les revues se regroupent en majorité dans les valeurs basses en termes de nombre de téléchargements par article. Les mesures de dispersion, quant à elles, indiquent que le nombre de téléchargements par article varie énormément d'une revue à l'autre, même au sein de chaque type de revue. Les valeurs tendent même à s'étendre de plus en plus avec le temps, en particulier pour les revues sans embargo.

Tableau 4. Statistiques d'usage basées sur le nombre moyen de téléchargements par article, pour les articles mis en ligne en 2011 (n = 1 124 508)

	Revues en libre accès différé					Revues en libre accès complet				
	2011	**2012**	**2013**	**2014**	**2015**	**2011**	**2012**	**2013**	**2014**	**2015**
Moyenne	27.02	48.29	94.20	150.32	166.79	127.12	254.04	257.95	260.17	230.80
Médiane	19.05	31.44	65.75	115.04	130.23	123.81	215.86	213.34	154.22	143.55
Étendue	144.11	266.45	396.20	576.55	496.08	310.73	600.82	638.90	788.45	822.97
Écart type	24.96	50.38	80.55	118.80	112.57	89.08	162.82	178.84	240.94	229.62
Coefficient de variation	92.40%	104.33%	85.51%	79.03%	67.49%	70.08%	64.09%	69.33%	92.61%	99.49%

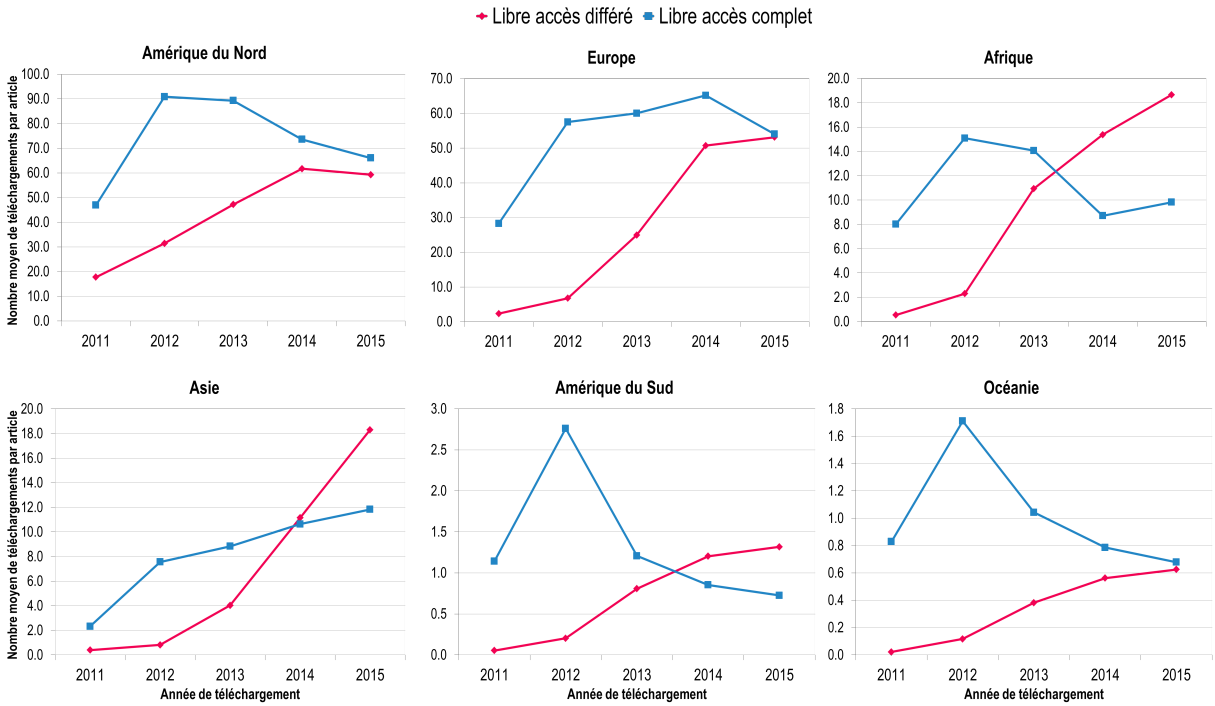

Figure 17. Nombre moyen de téléchargements par article pour les revues en libre accès différé et pour les revues en libre accès complet par continent. Articles mis en ligne en 2011 (n = 1 106 541)

La figure 17 montre la courbe d'usage par continent pour les articles ayant été mis en ligne en 2011 et dont l'embargo se termine en 2013. Encore une fois, les revues en libre accès différé sont mises en parallèle avec les revues en libre accès complet. Pour tous les continents, les revues en libre accès complet sont bien plus téléchargées que celles soumises à une période d'embargo. Celles-ci connaissent, dans tous les cas, une hausse importante de téléchargements dès la fin de l'embargo. La hausse est moins importante pour l'Océanie, mais cela peut s'expliquer par le faible nombre de téléchargements en provenance de ce continent. Si les courbes d'un continent à l'autre sont très différentes, elles tendraient probablement à se ressembler davantage avec un plus grand jeu de données (seulement 0,44% des téléchargements proviennent de l'Océanie, 1,03% de l'Amérique du Sud et 8,98% de l'Asie, comparativement à 37,11% pour l'Europe, 36,60% pour l'Amérique du Nord et 15,84% pour l'Afrique). Le passage au libre accès semble également moins marqué pour l'Amérique du Nord puisque bon nombre d'institutions canadiennes et américaines sont abonnées aux revues de la plateforme, mais nous verrons plus loin qu'il a tout de même un impact très positif. Les

tendances observées ci-dessus sont similaires pour le Canada, la France et les États-Unis, de même que pour le Québec et l'Ontario.

Le libre accès en synchronie

La figure 18 permet de confronter la proportion de téléchargements d'articles en libre accès effectués en 2015, pour chaque continent, à la proportion réelle d'articles libre accès dans la collection d'Érudit. Parmi les articles en libre accès de la collection, nous incluons les articles de revues en libre accès complet mis en ligne depuis 2011, ainsi que les articles pour lesquels nous avons pu calculer, au moyen des logs, que la période d'embargo était terminée, également mis en ligne depuis 2011. En 2015, 55,05% des articles disponibles dans la collection d'Érudit – qui avaient donc le potentiel d'être téléchargés – étaient en libre accès. Pourtant, dans tous les continents, la proportion d'articles en libre accès téléchargés est nettement supérieure à 55,05% : 67,24% des articles téléchargés sont en libre accès en Amérique du Nord, 85,08% en Europe, 82,97% en Asie, 86,53% en Afrique, 84,41% en Amérique du Sud et 83,16% en Océanie. La différence entre les deux proportions permet d'estimer l'avantage du libre accès pour chaque continent, et montre clairement que les articles en libre accès sont plus susceptibles d'être cités. L'Amérique du Nord, malgré le nombre d'institutions abonnées à la plateforme, télécharge également une proportion plus importante d'articles en libre accès que celle attendue.

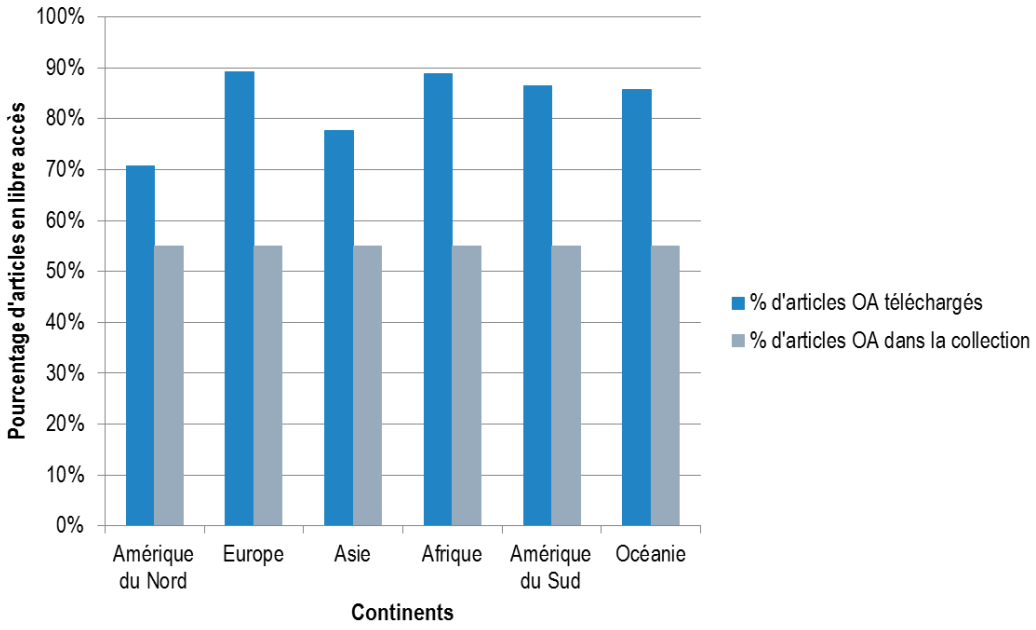

Figure 18. Pourcentage de téléchargements d'articles en libre accès par continent, comparé au pourcentage d'articles en libre accès dans la collection d'Érudit. Année de téléchargement fixée à 2015 ; articles mis en ligne entre 2011 et 2015 (n = 1 078 387)

Pour conclure, le passage au libre accès pour les revues soumises à une période d'embargo a un impact positif sur les téléchargements, en particulier pour les continents autres que l'Amérique du Nord. Les revues en libre accès complet, quant à elles, ont un avantage important en ce qui concerne le nombre moyen de téléchargements par article. De surcroît, si l'on considère l'ensemble des articles disponibles en libre accès à un moment donné, peu importe le type de revue dans lequel ils sont publiés, il apparaît que le libre accès a un impact positif sur les téléchargements.

Conclusion

Discussion

Dans le cadre de ce projet de recherche, nous avons cherché à contribuer à une meilleure compréhension de l'usage des revues nationales en sciences sociales et humaines. Plus précisément, l'usage des articles de revues savantes de la plateforme Érudit a été étudié de façon détaillée, les habitudes de téléchargement des usagers sur cette plateforme ont été décrites, et l'impact des politiques de libre accès des revues sur le nombre de téléchargements qu'elles reçoivent a été analysé. En parallèle, nous avions pour objectif d'éprouver les qualités des données de téléchargements.

Dans la première partie de nos résultats, une analyse exploratoire a permis de dresser le portrait général, et a révélé diverses informations sur la provenance des usagers, l'âge des articles téléchargés, les revues les plus utilisées, les différences entre les disciplines, l'utilisation du site Web d'Érudit par rapport à d'autres référents, les appareils et systèmes d'exploitation utilisés par les usagers et, enfin, l'évolution d'une année à l'autre pour toutes ces questions. Parmi les faits les plus intéressants à noter, nous avons constaté que l'âge des articles téléchargés (âge moyen de 16 ans et âge médian de 12 ans pour les sciences sociales et humaines (SSH), âge moyen de 11 ans et âge médian de 9 ans pour les sciences naturelles et génie (SNG)) sur la plateforme est en concordance avec les études de l'obsolescence de la littérature savante, qui ont démontré que les publications en SSH ont une plus grande durée de vie que celles en SNG. En outre, la majorité des téléchargements est, sans surprise, issue du Canada, de la France et des États-Unis, suivis par les pays d'Europe de l'Ouest et d'Afrique du Nord. Chose étonnante, près de 60% des téléchargements proviennent d'utilisateurs ayant utilisé l'un des services de Google, ce qui prouve l'importance pour le site d'Érudit d'être bien référencé sur le Web – sur ce sujet, rappelons le problème d'indexation qu'a connu la plateforme en 2014 et qui a eu un impact marqué sur le nombre de téléchargements.

Dans un second temps, notre analyse des habitudes de téléchargement des usagers par pays s'est inscrite dans la lignée des études sur la conciliation entre la vie personnelle et la vie professionnelle dans le monde académique. L'activité en ligne des usagers aux États-Unis

s'est révélée extrêmement différente de celle du Canada et de la France, peu importe l'échelle de temps observée (par heure de jour, par jour de la semaine, par mois et par saison). Les Américains, malgré que nous ayons tenu compte des différents fuseaux horaires et de l'utilisation de proxys, sont étonnamment actifs pendant la nuit, pendant les week-ends et même pendant l'été, ce qui est peut-être symptomatique de la grande pression exercée sur les chercheurs dans ce pays. Quoi qu'il en soit, nos résultats semblent montrer que la plupart des usagers au Canada sont des étudiants de 1er cycle universitaire, car ceux-ci sont plus actifs pendant l'année scolaire, mais bien moins pendant les Fêtes et la période estivale. En France, mais surtout aux États-Unis, les usagers paraissent plutôt être des chercheurs puisqu'ils utilisent Érudit en soirée, la nuit, le week-end, et même, en été. Le caractère particulier de l'emploi du temps des chercheurs américains a aussi été constaté par Wang et ses collaborateurs (2012) et, de surcroît, l'influence du calendrier et des événements spéciaux a été démontrée par Magnone (2013) et Moed et Halevi (2016). Cependant, les quelques études qui ont été menées sur cette question ne nous permettent pas de nous prononcer davantage sur les différences entre les pratiques des usagers de la plateforme selon leur provenance géographique.

En dernier lieu, nous avons été en mesure de confirmer que le passage au libre accès, pour les revues soumises à une période d'embargo, a un impact très positif sur les téléchargements, en particulier pour les continents autres que l'Amérique du Nord, où se trouve la majorité des abonnements à la plateforme. Les revues en libre accès complet ont un avantage évident par rapport aux revues en libre accès différé, dont les articles peinent à rattraper le retard qu'ils gagnent dans la période décisive suivant leur mise en ligne. De surcroît, lorsque l'on tient compte du nombre d'articles en accès restreint et en libre accès disponibles dans la collection d'Érudit à un moment précis dans le temps, on remarque que le nombre de téléchargements d'articles en libre accès est nettement supérieur au nombre de téléchargements d'articles en accès restreint, et ce, pour tous les continents. L'effet positif du libre accès est d'une importance cruciale pour la plateforme, favorable à la diffusion gratuite et en ligne des résultats de recherche. Les résultats que nous avons obtenus montrent toutefois les limites de la diffusion en libre accès différé, c'est-à-dire avec une période d'embargo.

Parmi les principales limites de notre étude, nous aimerions d'abord souligner que la partie exploratoire de nos résultats ne nous permet que de décrire les caractéristiques de l'usage des articles sur la plateforme Érudit, mais non pas d'expliquer les phénomènes observés ni de prédire les changements à venir. Ensuite, la partie quasi-expérimentale visant à mesurer l'impact du libre accès ne permet pas une aussi grande généralisabilité des résultats que ne l'aurait fait une véritable expérimentation, car les variables parasites sont plus difficiles à isoler dans une quasi-expérimentation. Mais pour des raisons évidentes, une expérimentation n'était pas envisageable pour notre étude (il nous aurait fallu un groupe contrôle d'articles en accès restreint, c.-à-d. dont l'embargo ne se terminerait jamais). Aussi, plus généralement, un téléchargement d'un article en texte intégral ne mène pas nécessairement à une lecture de l'article. Seule une étude qualitative pourrait permettre d'évaluer quelle est la proportion d'articles téléchargés qui sont lus en entier, en partie ou pas du tout. Enfin, le problème du calcul de la période d'embargo nous a contraints à nous limiter uniquement aux articles mis en ligne depuis 2011 pour tout ce qui concerne l'impact du libre accès sur l'usage des articles, alors que ces articles récents ne représentent qu'une petite portion de la collection.

Néanmoins, l'impact positif du libre accès sur l'usage des articles qui a été observé dans notre étude ne peut être attribué à aucun des biais reprochés dans d'autres études (p. ex. Kurtz et al., 2005a). En effet, le biais du « *early view access* » ne s'applique pas pour la plupart des articles qui deviennent disponibles en libre accès seulement après une période d'embargo. Le « *self-selection bias* » n'a pas non plus lieu d'être, car c'est la plateforme et non les auteurs qui détermine la politique de diffusion des articles. C'est donc bel et bien le postulat du « *OA advantage* » qui semble être en cause.

Limites des données de téléchargements

Dans le manifeste pour les altmetrics, plusieurs auteurs reprochaient aux indicateurs bibliométriques traditionnels tels que le nombre de citations d'être facilement manipulables et défendaient même l'idée que « mature altmetrics systems could be more robust, leveraging the diversity of of altmetrics and statistical power of big data to algorithmically detect and correct for fraudulent activity » (Priem, Taraborelli, Groth et Neylon, 2011). Contrairement à ces auteurs, notre projet a démontré que les téléchargements, qui font partie des nouvelles

métriques dites « alternatives », peuvent être aisément influencés par la manière dont est effectué le traitement des données et, surtout, par la manière dont sont exclus les robots (lorsqu'une technique de détection a été mise en place, ce qui est loin d'être toujours le cas dans la littérature). Pour cette raison, il n'est pas toujours possible de comparer les analyses des données de téléchargements entre différents jeux de données : « different publication archives that do generate such statistics [download statistics] may apply different ways to record and/or count downloads, meaning that results are not always directly comparable across archives » (Moed, 2012).

En dépit des biais qui peuvent être introduits au moment du nettoyage et du traitement des données, la plupart des études consultées dans notre revue de littérature ne décrit pas les choix qui ont été effectués en la matière. Quelques études utilisent toutefois un jeu de données provenant de PLOS Article-Level Metrics (ALM) (p. ex. Yan et Gerstein, 2011), qui a l'avantage d'être clairement décrit. En effet, PLOS utilise les standards COUNTER 3, décrit ses sources de données ainsi que leur traitement, donne la liste des robots exclus[87] et explique même les difficultés liées à l'interprétation des données d'usage :

> **Usage data should be interpreted with caution**. For example, **it should not be assumed that a certain amount of usage equates to a certain number of "real" people reading** (or even viewing) an article. Similarly, it may be impossible to absolutely compare the usage of one article against any other, due to the large number of factors that might be at play, many of which would not be obvious from the data provided. Therefore, while the understanding (and reporting) of usage data remains in its infancy, we recommend that you interpret these data with caution and **we suggest that you simply regard the numbers as an indicator of usage levels rather than an absolute measure of usage** (PLOS One, s. d.) [nous soulignons].

À notre avis, un exemple de bonne pratique est celui de l'étude de Priem, Piwowar et Hemminger (2012) qui est basée, entre autres, sur les données de PLOS ALM. La méthode est décrite de façon très rigoureuse et les auteurs donnent même accès à l'ensemble de leurs jeux de données et de leurs scripts en fournissant un lien vers GitHub (jasonpriem, s. d.). Il est d'ailleurs intéressant de signaler que quelques auteurs ont obtenu une corrélation entre le fait de donner accès aux données et le nombre de citations, notamment Henneken et Accomazzi

[87] L'emploi d'une liste standardisée est une technique plutôt rudimentaire de détection de robots, mais *PLOS* a le mérite de définir clairement les limites de son jeu de données. La transparence est, à notre avis, ce qui importe le plus, car, en ce qui concerne les données d'usage, il est impossible de parvenir à un nettoyage de données parfait.

(2011) dans le domaine de l'astronomie, ainsi que Piwowar, Day et Fridsma (2007) pour la médecine. En suivant l'exemple Priem, Piwowar et Hemminger (2012), nous avons décidé de rendre accessible sur GitHub le script de traitement de données qui a été développé par Yorrick Jansen, ingénieur en informatique, dans le cadre de notre projet (yorrick, s. d.).

La diffusion des résultats de recherche à l'ère de l'édition commerciale

Depuis quelques décennies, l'institution universitaire a connu d'importantes transformations liées à l'économie néolibérale des sociétés occidentales. Parmi ces bouleversements, Hall mentionne le développement d'un marché de l'éducation supérieure et, par le fait même, d'une concurrence féroce entre les institutions universitaires, l'augmentation du nombre d'étudiants et de la charge de travail administratif, la concentration de la recherche dans un petit nombre d'institutions prestigieuses, le déclin de la sécurité d'emploi et des conditions de travail des chercheurs, l'apparition de pratiques de microgestion pour contrôler et augmenter la productivité, le manque de soutien à certains champs d'études, en particulier ceux qui ne répondent pas à une logique marchande, la mutation des étudiants en clients, ainsi que l'exploitation des étudiants comme main-d'œuvre à bon marché (2008, p. 1-2). Or, l'une des fonctions essentielles de l'institution universitaire, depuis les premières *universitates* médiévales, est d'assurer l'accès à des sources d'information (Chodorow et Lyman, 1998, p. 61). Les universités ont donc une responsabilité de fournir l'information nécessaire à la recherche, mais aussi de préserver cet héritage intellectuel. Pour ce faire, il faut toutefois garder le contrôle sur l'information, et plusieurs considèrent que les universités de nos jours échouent à remplir cette mission (Chodorow et Lyman, 1998, p. 69-70).

La privatisation des statistiques d'utilisation (notamment les données de téléchargements) pourrait avoir des conséquences néfastes sur la publication savante, car celles-ci confèrent aux grands éditeurs commerciaux un pouvoir immense : « [i]l est quelque peu troublant de noter que de tels instruments, particulièrement puissants, sont en train d'être monopolisés par des intérêts privés […] [qui] peuvent surveiller, mesurer et peut-être même prédire » (Guédon, 2001, p. 41). Dans le même ordre d'idées, le militant Aaron Swartz mettait vivement en garde contre les pratiques de ces compagnies :

> Information is power. […] The world's entire scientific and cultural heritage, published over centuries in books and journals, is increasingly being digitized and locked up by a handful of private corporations. Want to read the papers featuring the most famous results of the sciences? You'll need to send enormous amounts to publishers like Reed Elsevier (Swartz, 2008).

Si l'importance de la plateforme Érudit pour la diffusion des résultats de recherche de langue française a été amplement décrite dans les chapitres précédents, rappelons que ce projet avait été mis sur pied dans le contexte d'oligopole des grands éditeurs commerciaux de revues savantes qui a vu le jour avec l'arrivée du numérique. Les résultats de notre étude ont montré que les revues de cette plateforme sont non seulement utilisées à l'échelle nationale, mais aussi à l'échelle internationale. Elle représente donc l'une des solutions au problème de couverture des SSH dans les grandes bases de données bibliographiques en valorisant les publications d'un intérêt plus local et celles dans une langue autre que l'anglais. Dans ce contexte, la durée actuelle de la barrière mobile nous apparaît comme un problème d'autant plus important. Nos résultats ont bien montré l'impact négatif de l'embargo sur l'usage des articles diffusés par Érudit, comme l'avait observé l'Institut des politiques publiques pour les revues françaises en SSH (Bacache-Beauvallet, Benhamou et Bourreau, 2015). Cela nous fait penser que la nouvelle politique des trois organismes subventionnaires canadiens sur le libre accès aux publications (Gouvernement du Canada, 2016), qui incite les titulaires d'une subvention à diffuser en libre accès leurs résultats de recherche dans les douze mois suivant la publication, ne peut constituer qu'une solution temporaire. Par conséquent, nous ne pouvons que recommander à Érudit la diminution, voire la suppression de la barrière mobile et le passage au libre accès complet pour l'ensemble des revues savantes, car cela augmenterait certainement l'usage et la pérennité des articles de la plateforme.

Bibliographie

Almind, T. C. et Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. *Journal of Documentation*, *53*(4), 404-426.

Antelman, K. (2004). Do open access articles have a greater research impact? *College & Research Libraries*, *65*(5), 372-382. doi:10.5860/crl.65.5.372

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Rebout, L. et Roberge, G. (2013). *Proportion of open access peer-reviewed papers at the European and world levels: 2004-2011* (p. 24). Montréal : Science-Metrix.

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V. et Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329-342.

Bacache-Beauvallet, M., Benhamou, F. et Bourreau, M. (2015). *Les revues de sciences humaines et sociales en France : libre accès et audience* (Rapport IPP n° 11). Repéré à http://www.ipp.eu/publication/juillet-2015-revues-sciences-humaines-et-sociales-shs-en-france-libre-acces-et-audience/

Beaudry, G., Boucher, M., Niemann, T. et Boismenu, G. (2009). Érudit : le numérique au service de l'édition en sciences humaines et sociales. *Mémoires du livre / Studies in Book Culture*, *1*(1). Repéré à http://id.erudit.org/iderudit/038637ar

Bergstrom, T. C., Courant, P. N., McAfee, R. P. et Williams, M. A. (2014). Evaluating big deal journal bundles. *Proceedings of the National Academy of Sciences*, *111*(26), 9425-9430. doi:10.1073/pnas.1403006111

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T. et Guðnason, G. (2010). Open access to the scientific journal literature: Situation 2009. *PLOS ONE*, *5*(6), e11273. doi:10.1371/journal.pone.0011273

Björneborn, L. (2004). *Small-world link structures across an academic Web space: A library and information science approach* (Thèse de doctorat, Royal School of Library and Information Science, Copenhagen, Denmark). Repéré à http://pure.iva.dk/ws/files/31034741/lennart_bjorneborn_phd.pdf

Björneborn, L. et Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, *55*(14), 1216-1227. doi:10.1002/asi.20077

Boismenu, G. et Beaudry, G. (2002). *Le nouveau monde numérique : le cas des revues universitaires*. Montréal : Presses de l'Université de Montréal.

Bornmann, L. (2012). Measuring the societal impact of research. *EMBO Reports*, *13*(8), 673-676. doi:10.1038/embor.2012.99

Bourdieu, P. (1976). Le champ scientifique. *Actes de la recherche en sciences sociales*, *2*(2), 88-104. doi:10.3406/arss.1976.3454

Brody, T., Carr, L. et Harnad, S. (2002). Evidence of hypertext in the scholarly archive. Dans *HYPERTEXT '02 Proceedings of the thirteenth ACM conference on hypertext and hypermedia* (p. 74-75). New York : ACM. doi:10.1145/513338.513359

Brody, T., Harnad, S. et Carr, L. (2006). Earlier Web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*(8), 1060-1072. doi:10.1002/asi.20373

Brookes, B. C. (1990). Biblio-, sciento-, infor-metrics?? What are we talking about ? Dans Rousseau, R. et Egghe, L. (dir.), *Informetrics 89/90: Selection of papers submitted for the second International Conference on Bibliometrics, Scientometrics, and Informetrics, London, Ontario, Canada, 5-7 July 1989*, Amsterdam : Elsevier. Repéré à https://uhdspace.uhasselt.be/dspace/handle/1942/857

Burton, R. E. et Kebler, R. W. (1960). The « half-life » of some scientific and technical literatures. *American Documentation*, *11*(1), 18-22. doi:10.1002/asi.5090110105

Cabanac, G. et Hartley, J. (2013). Issues of work-life balance among *JASIST* authors and editors. *Journal of the American Society for Information Science and Technology*, *64*(10), 2182-2186. doi:10.1002/asi.22888

Calver, M. C. et Bradley, J. S. (2010). Patterns of citations of open access and non-open access conservation biology journal papers and book chapters. *Conservation Biology*, *24*(3), 872-880. doi:10.1111/j.1523-1739.2010.01509.x

Caza, P.-É. (2014). L'avenir numérique des revues savantes. *Actualités UQAM*. Repéré à http://www.actualites.uqam.ca/2014/avenir-numerique-des-revues-savantes

Chevallier, S. et Chauviré, C. (2010). *Dictionnaire Bourdieu*. Paris : Ellipses.

Chi, P.-S. (2016). Differing disciplinary citation concentration patterns of book and journal literature? *Journal of Informetrics*, *10*(3), 814-829. doi:10.1016/j.joi.2016.05.005

Chi, P.-S., Jeuris, W., Thijs, B. et Glänzel, W. (2015). Book bibliometrics: A new perspective and challenge in indicator building based on the Book Citation Index. Dans *Proceedings of ISSI 2015: The 15th International Conference on Scientometrics and Informetrics* (p. 1161-1169). Repéré à https://lirias.kuleuven.be/handle/123456789/507822

Chodorow, S. et Lyman, P. (1998). The responsabilities of universities in the new information environment. Dans *The mirage of continuity: Reconfiguring academic information resources for the 21st century* (p. 61-104). Repéré à http://web.library.emory.edu/FryeInstitute/Readings/16122513.pdf

Cloutier, M. (2015). *La publication en libre accès : étude sur les niveaux de connaissance et d'intérêt de la communauté de recherche des Facultés d'administration, de droit, d'éducation, d'éducation physique et sportive, de lettres et sciences humaines et de théologie et d'études religieuses de l'Université de Sherbrooke* (Mémoire de maîtrise, Université de Sherbrooke). Repéré à http://savoirs.usherbrooke.ca/handle/11143/7940

Commission européenne. (2012). *Recommandation de la Commission du 17.7.2012 relative à l'accès aux informations scientifiques et à leur conservation* (p. 10). Bruxelles. Repéré à http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_fr.pdf

Contat, O. et Gremillet, A.-S. (2015). Publier : à quel prix ? Étude sur la structuration des coûts de publication pour les revues françaises en SHS. *Revue française des sciences de l'information et de la communication*, (7). doi:10.4000/rfsic.1716

Couture, M. (2013). L'accès libre aux publications de recherche. Repéré à http://benhur.teluq.uquebec.ca/~mcouture/oa/accesLibre.htm

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A. et Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science*, *49*(14), 1319-1328. doi:10.1002/(SICI)1097-4571(1998)49:14<1319::AID-ASI9>3.0.CO;2-W

Crowther, A. (1999). Consortia licensing, information as infrastructure. Dans *Proceedings of the IATUL Conferences* (p. 9). Repéré à http://docs.lib.purdue.edu/iatul/1998/papers/9

Davis, P. M. (2011a). Do discounted journal access programs help researchers in sub-Saharan Africa? A bibliometric analysis. *Learned Publishing*, *24*(4), 287-298.

Davis, P. M. (2011b). Open access, readership, citations: A randomized controlled trial of scientific journal publishing. *The FASEB Journal*, *25*(7), 2129-2134. doi:10.1096/fj.11-183988

Davis, P. M. et Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, *71*(2), 203-215. doi:10.1007/s11192-007-1661-8

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G. et Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: Randomised controlled trial. *BMJ*, *337*, 1-6. doi:10.1136/bmj.a568

Delamont, S. (1989). Citation and social mobility research: Self defeating behaviour? *The Sociological Review*, *37*(2), 332-337. doi:10.1111/j.1467-954X.1989.tb00032.x

Diodato, V. P. (1994). *Dictionary of bibliometrics*. New York : Haworth Press.

Dobrov, G. M. et Korennoi, A. A. (1969). The informational basis of scientometrics. Dans Fédération internationale de documentation, *On theoretical problems of informatics* (p. 165-191). Moscow : All-Union Institute for Scientific and Technical Information.

Doran, D. et Gokhale, S. S. (2010). Web robot detection techniques: Overview and limitations. *Data Mining and Knowledge Discovery*, *22*(1-2), 183-210. doi:10.1007/s10618-010-0180-z

Érudit. (2015). Rapport annuel 2014-2015. Repéré à http://erudit.org/rapport/2015/

Érudit. (2016). The electronic text centre at UNB libraries. Repéré à http://www.erudit.org/revue/?mode=fond&idfond=3&lettre

Érudit. (s.d.-a). Mission. Repéré à https://apropos.erudit.org/fr/erudit/mission/

Érudit. (s.d.-b). Politique d'Érudit sur le libre accès aux publications de la recherche. Repéré à http://www.erudit.org/documents/apropos/Eruditla.pdf

Evans, J. A. et Reimer, J. (2009). Open access and global participation in science. *Science*, *323*(5917), 1025-1026. doi:10.1126/science.1154562

Eve, M. P. (2014). *Open access and the humanities: Contexts, controversies and the future*. Cambridge : Cambridge University Press. Repéré à http://dx.doi.org/10.1017/CBO9781316161012

Eysenbach, G. (2006). Citation advantage of open access articles. *PLOS Biology*, *4*(5), e157, 0692-0698. doi:10.1371/journal.pbio.0040157

Eysenbach, G. (2008). Word is still out: Publication was premature. Repéré à http://www.bmj.com/rapid-response/2011/11/02/word-still-out-publication-was-premature

Fondation canadienne pour l'innovation. (2017). Fonds des initiatives scientifiques majeures. Repéré à https://www.innovation.ca/fr/le-financement/fonds-des-initiatives-scientifiques-majeures

Fortin, F. et Gagnon, J. (2010). *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives* (2e éd.). Montréal : Chenelière éducation.

Fournier, M., Germain, A., Lamarche, Y. et Maheu, L. (1975). Le champ scientifique québécois : structure, fonctionnement et fonctions. *Sociologie et sociétés*, *7*(1), 119-132.

Fournier, M. et Maheu, L. (1975). Nationalismes et nationalisation du champ scientifique québécois. *Sociologie et sociétés*, *7*(2), 89-114.

Frandsen, T. F. (2009). Attracted to open access journals: A bibliometric author analysis in the field of biology. *Journal of Documentation*, *65*(1), 58-82. doi:10.1108/00220410910926121

Frazier, K. (2001). The librarians' dilemma: Contemplating the costs of the « big deal ». *D-Lib Magazine*, *7*(3). doi:10.1045/march2001-frazier

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108-111. doi:10.1126/science.122.3159.108

Garfield, E. (1983). *Citation indexing: Its theory and application in science, technology, and humanities*. Philadelphia : ISI Press.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T. et Harnad, S. (2010). Self-Selected or mandated, open access increases citation impact for higher quality research. *PLOS ONE*, *5*(10), e13636. doi:10.1371/journal.pone.0013636

Gargouri, Y., Larivière, V., Gingras, Y., Brody, T., Carr, L. et Harnad, S. (2012). Testing the Finch hypothesis on green OA mandate effectiveness. *arXiv:1210.8174 [cs]*. Repéré à http://arxiv.org/abs/1210.8174

Gargouri, Y., Larivière, V., Gingras, Y., Carr, L. et Harnad, S. (2012). Green and gold open access percentages and growth, by discipline. *arXiv:1206.3664 [cs]*. Repéré à http://arxiv.org/abs/1206.3664

Garvey, W. D. et Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist*, *26*(4), 349-362. doi:10.1037/h0032059

Gaulé, P. (2009). Access to scientific literature in India. *Journal of the American Society for Information Science and Technology*, *60*(12), 2548-2553. doi:10.1002/asi.21195

Gaulé, P. et Maystre, N. (2011). Getting cited: Does open access help? *Research Policy*, *40*(10), 1332-1338. doi:10.1016/j.respol.2011.05.025

Geens, N., Huysmans, J. et Vanthienen, J. (2006). Evaluation of Web robot discovery techniques: A benchmarking study. Dans P. Perner (dir.), *Advances in data mining: Applications in medicine, Web mining, marketing, image and signal mining* (p. 121-130). Springer : Berlin Heidelberg. doi:10.1007/11790853_10

Gentil-Beccot, A., Mele, S. et Brooks, T. (2010). Citing and reading behaviours in high-energy physics. *Scientometrics*, *84*(2), 345-355. doi:10.1007/s11192-009-0111-1

Gingras, Y. (1984). La valeur d'une langue dans un champ scientifique. *Recherches sociographiques*, *25*(2), 285. doi:10.7202/056095ar

Gingras, Y. (2008). Du mauvais usage de faux indicateurs. *Revue d'histoire moderne et contemporaine*, *n° 55-4bis*(5), 67-79. Repéré à http://www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2008-5-page-67.htm

Gingras, Y. (2014). *Les dérives de l'évaluation de la recherche : du bon usage de la bibliométrie*. Paris : Éditions Raisons d'Agir.

Gingras, Y. (2015, 9 septembre). *La dérive des facteurs d'impact pour les revues savantes* Repéré à https://apropos.erudit.org/fr/la-derive-des-facteurs-dimpact-pour-les-revues-savantes/

Gingras, Y. et Mosbah-Natanson. (2010). Where are social sciences produced? Dans UNESCO et International Social Science Council (dir.), *World social science report: knowledge divides* (p. 149-153). Paris : UNESCO Publishing. Repéré à http://www.unesco.org/new/en/social-and-human-sciences/resources/reports/world-social-science-report

Ginsparg, P. (2011). ArXiv at 20. *Nature*, *476*(7359), 145-147. doi:10.1038/476145a

Glänzel, W. et Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37-53. doi:10.1177/016555159502100104

Glänzel, W. et Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, *35*(1), 31-44. doi:10.1016/S0306-4573(98)00028-4

Gosnell, C. (1944). Obsolescence of books in college libraries. *College & Research Libraries*, *5*(1), 115-125.

Gouvernement du Canada. (2016). Politique des trois organismes sur le libre accès aux publications. Repéré à http://www.science.gc.ca/default.asp?lang=Fr&n=F6765465-1

Greyson, D., Morgan, S., Hanley, G. et Wahyuni, D. (2009). Open access archiving and article citations within health services and policy research. *Journal of the Canadian Health Libraries Association (JCHLA) / Journal de l'Association des bibliothèques de la santé du Canada (JABSC)*, *30*(2), 51-58.

Guare, J. (1990). *Six degrees of separation: A play*. New York : Vintage Books. Repéré à https://books.google.com/books?hl=fr&lr=&id=SHsyHb5s5xgC&oi=fnd&pg=PA3&dq=guare+Six+Degrees+of+Separation&ots=Uza-8AZd-T&sig=ej_c8NhgCWhEKPyDmjTF8RJuiU8

Guédon, J.-C. (2001). À l'ombre d'Oldenburg : bibliothécaires, chercheurs scientifiques, maisons d'édition et le contrôle des publications scientifiques. *ARL Meeting, Toronto, mai 2001*. Repéré à https://halshs.archives-ouvertes.fr/halshs-00395366/document

Guest, D. E. (2002). Perspectives on the study of work-life balance. *Social Science Information*, *41*(2), 255-279. doi:10.1177/0539018402041002005

Habib, A. (2010). Challenging the international academic publishing industry. Dans UNESCO et International Social Science Council (dir.), *World social science report: Knowledge divides* (p. 311-313). Paris : UNESCO Publishing. Repéré à http://www.unesco.org/new/en/social-and-human-sciences/resources/reports/world-social-science-report

Hajjem, C., Harnad, S. et Gingras, Y. (2005). Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. Repéré à http://arxiv.org/abs/cs/0606079

Hall, G. (2008). *Digitize this book! The politics of new media, or why we need open access now*. Minneapolis : University of Minnesota Press.

Harmon, J. E. et Gross, A. G. (2007). *The scientific literature: A guided tour*. Chicago et London : The University of Chicago Press.

Harnad, S. (2008). Davis et al's 1-year study of self-selection bias: No self-archiving control, no OA effect, no conclusion. Repéré à http://www.bmj.com/rapid-response/2011/11/02/davis-et-als-1-year-study-self-selection-bias-no-self-archiving-control-no

Harnad, S. (2011). Open access to research: Changing researcher behavior through university and funder mandates. *JeDEM Journal of Democracy and Open Government*, *3*(1), 33-41.

Harnad, S. et Brody, T. (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, *10*(6). Repéré à http://www.dlib.org/dlib/june04/harnad/06harnad.html

Harnad, S. et Carr, L. (2000). Integrating, navigating and analyzing eprint archives through Open Citation-linking (the OpCit Project). *Current Science*, *79*(5), 629-638.

Haustein, S. (2014). Readerschip metrics. Dans B. Cronin et C. R. Sugimoto (dir.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (p. 327-344). Cambridge, Mass. et London : The MIT Press.

Haustein, S., Bowman, T. D. et Costas, R. (2016). Interpreting « altmetrics »: Viewing acts on social media through the lens of citation and social theories. Dans C. R. Sugimoto (dir.), *Theories of informetrics and scholarly communication* (p. 372-405). Berlin : De Gruyter.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D. et Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *Information Technology*, *56*(5). doi:10.1515/itit-2014-1048

Henneken, E. A. et Accomazzi, A. (2011). Linking to data: Effect on citation rates in astronomy. *arXiv:1111.3618 [astro-ph]*. Repéré à http://arxiv.org/abs/1111.3618

Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D. et Murray, S. S. (2006). Effect of e-printing on citation rates in astronomy and physics. *arXiv:cs/0604061*. Repéré à http://arxiv.org/abs/cs/0604061

Henry, G. (2015, 10 avril). Investissement de 1,4 million $ dans la plateforme Érudit par la Fondation canadienne pour l'innovation [Érudit : le blogue]. Repéré à https://apropos.erudit.org/fr/investissement-de-14-million-dans-la-plateforme-erudit-par-la-fondation-canadienne-pour-linnovation/

Henry, G. (2016, 25 novembre). Succès grandissant pour le corpus d'Érudit aux États-Unis [Érudit : le blogue]. Repéré à https://apropos.erudit.org/fr/succes-grandissant-pour-le-corpus-derudit-aux-etats-unis/

Hickman, I. (2000). Mining the social life of an eprint archive. Repéré à http://opcit.eprints.org/ijh198/index.html

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193-215. doi:10.1007/BF02457380

Hitchcock, S. (s.d.). The effect of open access and downloads ('hits') on citation impact: A bibliography of studies. *The Open Citation Project*. Repéré à http://opcit.eprints.org/oacitation-biblio.html

Hitchcock, S., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., … Harnad, S. (2002). Open citation linking: the way forward. *D-Lib Magazine*, *8*(10). doi:10.1045/october2002-hitchcock

Hitchcock, S., Brody, T., Gutteridge, C., Carr, L. et Harnad, S. (2003). The impact of OAI-based search on access to research journal papers. *Serials*, *16*(3), 255-260.

Houghton, B. (1975). *Scientific periodicals: Their historical development, characteristics and control*. London : Clive Bingley.

Houghton, J. (2009). *Open access: What are the economic benefits? A comparison of the United Kingdom, Netherlands and Denmark* (Rapport Knowledge Exchange). Victoria University : Melbourne.

Humanité. (s. d.). *Trésor de la langue française informatisé*. Repéré à http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?29;s=1938573495;r=2;nat=;sol=1

Huntington, P., Nicholas, D., Jamali, H. R. et Tenopir, C. (2006). Article decay in the digital environment: An analysis of usage of OhioLINK by date of publication, employing deep log methods. *Journal of the American Society for Information Science and Technology*, *57*(13), 1840-1851. doi:10.1002/asi.20383

IDATE et Cairn.info. (2015a). L'open access et les revues SHS de langue française : tendances du secteur, environnement réglementaire et perspectives 2018. Repéré à http://www.openaccess-shs.info/wp-content/uploads/2015/10/Etude-IDATE-CAIRN-INFO-20151002.pdf

IDATE et Cairn.info. (2015b). L'open access et les revues SHS de langue française. Repéré à http://www.openaccess-shs.info/lopen-access-et-les-revues-shs-de-langue-francaise/

Ingwersen, P. et Björneborn, L. (2004). Methodological issues of webometric studies. Dans H. Moed, W. Glänzel et U. Schmoch (dir.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies on S&T systems* (p. 339-369). Dordrecht : Kluwer Academic Publishers.

IP Location. (2016). Where is geolocation of an IP address? Repéré à https://www.iplocation.net/

Jahandideh, S., Abdolmaleki, P. et Asadabadi, E. B. (2007). Prediction of future citations of a research paper from number of its internet downloads. *Medical Hypotheses*, *69*(2), 458-459. doi:10.1016/j.mehy.2007.01.007

jasonpriem. (2010). I like the term #articlelevelmetrics, but it fails to imply *diversity* of measures. Lately, I'm liking #altmetrics. Repéré à https://twitter.com/jasonpriem/status/25844968813

jasonpriem. (s. d.). plos_altmetrics_study. Repéré à https://github.com/jasonpriem/plos_altmetrics_study

Jonkers, K. (2010). The share of major social science disciplines in bibliometric databases. Dans UNESCO et International Social Science Council (dir.), *World social science report: Knowledge divides* (p. 194-196). Paris : UNESCO Publishing. Repéré à http://www.unesco.org/new/en/social-and-human-sciences/resources/reports/world-social-science-report

Kaufman, P. (1998). Structure and crisis: Markets and market segmentation in scholarly publishing. Dans *The mirage of continuity: Reconfiguring academic information resources for the 21st century* (p. 178-192). Repéré à http://web.library.emory.edu/FryeInstitute/Readings/16122513.pdf

Kousha, K. et Thelwall, M. (2015). Web indicators for research evaluation. Part 3: Books and non-standard outputs. *Profesional De La Informacion*, *24*(6), 724-736. doi:10.3145/epi.2015.nov.04

Kuhn, T. S. (1983). *La structure des révolutions scientifiques* (traduit par L. Meyer). Paris : Flammarion.

Kurtz, M. J. et Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology*, *44*(1), 1-64. doi:10.1002/aris.2010.1440440108

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E. A. et Murray, S. S. (2005a). The effect of use and access on citations. *Information Processing & Management*, *41*(6), 1395-1402. doi:10.1016/j.ipm.2005.03.010

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M. et Murray, S. S. (2005b). Worldwide use and impact of the NASA Astrophysics Data System digital library. *Journal of the American Society for Information Science and Technology*, *56*(1), 36-45. doi:10.1002/asi.20095

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., … Elwell, B. (2005c). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, *56*(2), 111-128. doi:10.1002/asi.20096

Kurtz, M. J. et Henneken, E. A. (2016). Measuring metrics: A 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. *Journal of the Association for Information Science and Technology*, 1-14. doi:10.1002/asi.23689

Lansingh, V. C. et Carter, M. J. (2009). Does open access in ophthalmology affect how articles are subsequently cited in research? *Ophthalmology*, *116*(8), 1425-1431. doi:10.1016/j.ophtha.2008.12.052

Larivière, V. (2014). De l'importance des revues de recherche nationales. *Découvrir*. Repéré à http://www.acfas.ca/publications/decouvrir/2014/09/l-importance-revues-recherche-nationales

Larivière, V., Archambault, É., Gingras, Y. et Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, *57*(8), 997–1004.

Larivière, V., Gingras, Y. et Archambault, É. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288-296. doi:10.1002/asi.20744

Larivière, V., Haustein, S. et Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLOS ONE*, *10*(6), e0127502. doi:10.1371/journal.pone.0127502

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*. Repéré à http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

Lebel, J. (2016). Revues savantes francophones au Canada : un premier portrait. *Découvrir*. Repéré à http://www.acfas.ca/publications/decouvrir/2016/06/revues-savantes-francophones-canada-premier-portrait

Li, X., Thelwall, M. et Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*(2), 461-471. doi:10.1007/s11192-011-0580-x

Library of Congress. (1994). *CONSER editing guide*. Washington : Serial Record Division, Library of Congress, Cataloging Distribution Service.

Line, M. B. (1993). Changes in the use of literature with time: Obsolescence revisited. *Library Trends*, *41*(4), 665-683.

Lippi, G. et Favaloro, E. J. (2012). Article downloads and citations: Is there any relationship? *Clinica Chimica Acta*, *415*, 195. doi:10.1016/j.cca.2012.10.037

Lorentzen, D. G. (2014). Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. *Scientometrics*, *99*(2), 409-445. doi:10.1007/s11192-013-1227-x

Magnone, E. (2013). A scientometric look at calendar events. *Journal of Informetrics*, *7*, 101-108. doi:10.1016/j.joi.2012.09.006

Manten, A. A. (1980). The growth of European scientific journal publishing before 1850. Dans A. J. Meadows (dir.), *Development of science publishing in Europe* (p. 1-22). Amsterdam, New York et Oxford : Elsevier Science.

McCabe, M. J. et Snyder, C. M. (2011). Did online access to journals change the economics literature? *Social Science Research Network (SSRN), 23*. Repéré à http://idei.fr/sites/default/files/medias/doc/conf/sic/mccabe.pdf

Meadows, A. J. (1974). *Communication in science*. London : Butterworths.

Meadows, A. J. (1980). Access to the results of scientific research: Developments in Victorian Britain. Dans A. J. Meadows (dir.), *Development of science publishing in Europe* (p. 43-62). Amsterdam, New York et Oxford : Elsevier Science.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago : University of Chicago Press.

Metcalfe, T. S. (2005). The Rise and Citation Impact of astro-ph in Major Journals. *arXiv:astro-ph/0503519*. Repéré à http://arxiv.org/abs/astro-ph/0503519

Metcalfe, T. S. (2006). The citation impact of digital preprint archives for solar physics papers. *Solar Physics*, *239*(1-2), 549-553. doi:10.1007/s11207-006-0262-7

Milgram, S. (1967). The small world problem. *Psychology Today*, *2*(1), 60-67.

Mingers, J. et Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1-19.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, *56*(10), 1088-1097. doi:10.1002/asi.20200

Moed, H. F. (2007). The effect of « open access » on citation impact: An analysis of arXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, *58*(13), 2047-2054. doi:10.1002/asi.20663

Moed, H. F. et Halevi, G. (2016). On full text download and citation distributions in scientific-scholarly journals. *Journal of the Association for Information Science and Technology*, *67*(2), 412-431. doi:10.1002/asi

Mosbah-Natanson, S. et Gingras, Y. (2014). The globalization of social sciences? Evidence from a quantitative analysis of 30 years of production, collaboration and citations in the social sciences (1980-2009). *Current Sociology*, *62*(5), 626-646. doi:10.1177/0011392113498866

Mounce, R. (2015). Another day, another Elsevier website illegally selling articles. *The Winnower*. Repéré à https://thewinnower.com/papers/another-day-another-elsevier-website-illegally-selling-articles

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: a review. *Scientometrics*, *66*(1), 81-100. doi:10.1007/s11192-006-0007-2

Nicholas, D., Huntington, P., Dobrowolski, T., Rowlands, I., Jamali M., H. R. et Polydoratou, P. (2005). Revisiting 'obsolescence' and journal article 'decay' through usage data: An analysis of digital journal use by year of publication. *Information Processing & Management*, *41*(6), 1441-1461. doi:10.1016/j.ipm.2005.03.014

Nicholas, D. et Rowlands, I. (2005). *New journal publishing models: An international survey of senior researchers* (A CIBER report for the Publishers Association and the International Association of STM Publishers) (p. 75). Repéré à http://www.homepages.ucl.ac.uk/~uczciro/pa_stm_final_report.pdf

Norris, M., Oppenheim, C. et Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, *59*(12), 1963-1972. doi:10.1002/asi.20898

Okerson, A. S. et O'Donnell, J. J. (dir.). (1995). *Scholarly journals at the crossroads: A subversive proposal for electronic publishing*. Washington, D.C. : Office of Scientific & Academic Publishing / Association of Research Libraries. Repéré à http://hdl.handle.net/2027/mdp.39015034923758

O'Leary, D. E. (2008). The relationship between citations and number of downloads in Decision Support Systems. *Decision Support Systems*, *45*(4), 972-980. doi:10.1016/j.dss.2008.03.008

Parker, R. H. (1982). Bibliometric models for management of an information store. II. Use as a function of age of material. *Journal of the American Society for Information Science*, *33*(3), 129-133. doi:10.1002/asi.4630330304

Perneger, T. V. (2004). Relation between online « hit counts » and subsequent citations: Prospective study of research papers in the BMJ. *British Medical Journal*, *329*(7465), 546-547. doi:10.1136/bmj.329.7465.546

Piwowar, H. A., Day, R. S. et Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLOS ONE*, *2*(3), e308. doi:10.1371/journal.pone.0000308

PLOS ONE. (s. d.). Usage data help. Repéré à http://journals.plos.org/plosone/s/usage-data-help

Price, D. J. de S. (1970). Citation measures of hard science, soft science, technology, and non-science. Dans *Communication among scientists and engineers* (p. 1-12). Lexington, Mass.: Health Lexington Books.

Price, D. J. de S. (1979). Networks of scientific papers. Dans A. J. Meadows (dir.), *The scientific journal* (p. 157-162). London : Aslib.

Priem, J. (2014). Altmetrics. Dans B. Cronin et C. R. Sugimoto (dir.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (p. 263-287). Cambridge, Mass. et London : The MIT Press.

Priem, J. et Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7). Repéré à http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2874

Priem, J., Piwowar, H. A. et Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv:1203.4745 [cs.DL]*. Repéré à http://arxiv.org/abs/1203.4745

Priem, J., Taraborelli, D., Groth, P. et Neylon, C. (2011). Altmetrics: A manifesto. Repéré à http://altmetrics.org/manifesto/

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, *25*, 348-349.

Riera, M. et Aibar, E. (2013). ¿Favorece la publicación en abierto el impacto de los artículos científicos? Un estudio empírico en el ámbito de la medicina intensiva. *Medicina Intensiva*, *37*(4), 232-240. doi:10.1016/j.medin.2012.04.002

Rousseau, R. (1988). Citation distribution of pure mathematics journals. Dans *Informetrics 87/88: Select proceedings of the first international conference on bibliometrics and theoretical aspects of information retrieval* (p. 249-260). Amsterdam : Elsevier Science. Repéré à https://doclib.uhasselt.be/dspace/bitstream/1942/844/1/rousseau249.pdf

Schwarz, G. J. et Kennicutt Jr, R. C. (2004). Demographic and citation trends in astrophysical journal papers and preprints. *arXiv:astro-ph/0411275*. Repéré à http://arxiv.org/abs/astro-ph/0411275

Science. (s. d.). *Trésor de la langue française informatisé*. Repéré à http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?90;s=1938573495;r=4;nat=;sol=1

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9). Repéré à http://search.proquest.com/docview/1301253581/citation/85A1A45CAEEC4ECBPQ/1

Sivertsen, G. et Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, *91*(2), 567-575. doi:10.1007/s11192-011-0615-3

Storer, N. W. (1967). The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association*, *55*, 75-84.

Suber, P. (2012). *Open access*. Cambridge, Mass. : MIT Press.

Swartz, A. (2008). Guerilla open access manifesto. Repéré à https://archive.org/stream/GuerillaOpenAccessManifesto/Goamjuly2008_djvu.txt

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, *28*(1), 1-3. doi:10.1016/0306-4573(92)90087-G

Thackray, A. et Brock, D. C. (2000). Eugene Garfield : History, scientific information and chemical endeavour. Dans B. Cronin, E. Garfield et H. B. Atkins (dir.), *The Web of knowledge: A festschrift in honor of Eugene Garfield* (p. 11-22). Medford, N. J. : Information Today. Repéré à https://books.google.ca/books?id=8O1kw0S6iLsC&pg=PR5&hl=fr&source=gbs_selected_pages&cad=3#v=onepage&q&f=false

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science*, *34*(4), 605-621. doi:10.1177/0165551507087238

Thelwall, M. (2009). *Introduction to webometrics: Quantitative Web research for the social sciences.* New York : Morgan & Claypool.

Thelwall, M., Vaughan, L. et Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, *39*(1), 81-135. doi:10.1002/aris.1440390110

Tijssen, R., Visser, M. et van Leeuwen, T. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, *54*(3), 381-397. doi:10.1023/A:1016082432660Vézina, K. (2006). Libre accès à la recherche scientifique : opinions et pratiques des chercheurs au Québec. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *1*(1). Repéré à https://journal.lib.uoguelph.ca/index.php/perj/article/view/103

Wang, X., Liu, C., Mao, W. et Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics*, *103*(2), 555-564. doi:10.1007/s11192-015-1547-0

Wang, X., Xu, S., Peng, L., Wang, Z., Wang, C., Zhang, C. et Wang, X. (2012). Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics*, *6*(4), 655-660. doi:10.1016/j.joi.2012.07.003

Whitehead, M. et Owen, B. (2016). *Universités canadiennes et édition pérenne : un livre blanc*. Ottawa : Association des bibliothèques de recherche du Canada. Repéré à http://www.carl-abrc.ca/wp-content/uploads/2016/04/Univ_can_editions_perennes_2016.pdf

Yan, K.-K. et Gerstein, M. (2011). The spread of scientific information: Insights from the Web usage statistics in *PLOS* article-level metrics. *PLOS ONE*, *6*(5), e19917. doi:10.1371/journal.pone.0019917

yorrick. (s. d.). download-data. Repéré à https://github.com/yorrick/download-data

Annexe 1 : Disciplines couvertes dans la plateforme Érudit (classification du National Science Foundation)

Grandes disciplines	Disciplines	Spécialités	Nombre de revues
Sciences sociales et humaines	**Humanités**	Histoire	10
		Humanités divers	9
		Littérature	8
		Langue et linguistique	6
		Religion	3
		Philosophie	2
	Sciences sociales	Sociologie	8
		Anthropologie et archéologie	5
		Économie	3
		Géographie	3
		Sciences politiques et administration publiqu	3
		Sciences sociales générales	3
		Sciences sociales divers	2
		Relations internationales	1
		Criminologie	1
		Démographie	1
		Étude sur la science	1
	Champs professionnels	Éducation	7
		Travail social	5
		Droit	4
		Management	3
		Bibliothéconomie et archivistique	1
		Champs professionnels divers	1
	Arts	Arts du spectacle	4
		Beaux-arts et architecture	2
	Santé	Gériatrie et gérontologie	1
		Réhabilitation	1
	Psychologie	Psychanalyse	1
Sciences naturelles et génie	**Sciences de la terre et de l'espace**	Science environnementale	3
	Médecine clinique	Psychiatrie	2
	Biologie	Botanique	1
	Recherche biomédicale	Recherche biomédicale - général	1

Annexe 2 : Informations contenues dans les logs et informations calculées lors du traitement des données

Principales données	Exemple	Source
Identifiant unique de l'article donné par Érudit et que l'on trouve dans l'URL du téléchargement	68814	Dans les logs
Moment du téléchargement selon l'heure de Montréal	1/5/2011 11:07	Dans les logs
Moment du téléchargement selon l'heure locale où le téléchargement a eu lieu	1/5/2011 17:07	Calculé lors du traitement des données
Adresse IP du proxy s'il y a lieu	132.204.2.131	Dans les logs
Adresse IP de l'usager	93.19.1.24	Dans les logs
Référent abrégé	Google	Calculé lors du traitement des données
Continent du téléchargement	EU	Calculé lors du traitement des données
Pays du téléchargement	France	Calculé lors du traitement des données
Ville du téléchargement	Europe/Paris	Calculé lors du traitement des données
Fuseau horaire	Paris	Calculé lors du traitement des données
Navigateur utilisé	Firefox	Calculé lors du traitement des données
Système d'exploitation de l'usager	Windows XP	Calculé lors du traitement des données
Appareil utilisé par l'usager (p = pc, t = tablet, m = mobile)	p	Calculé lors du traitement des données
Âge de l'article téléchargé (année du téléchargement - année de publication de l'article)	41	Calculé lors du traitement des données
L'article téléchargé est-il sous embargo? (vrai ou faux)	f	Calculé lors de la création de la base de données
Abréviation du nom de la revue utilisé dans l'URL	haf	Dans les logs
Discipline générale selon la classification du NSF	Sciences sociales et humaines	Calculé lors du traitement des données
Discipline selon la classification du NSF	Humanités	Calculé lors du traitement des données
Spécialité selon la classification du NSF	Histoire	Calculé lors du traitement des données
Année de publication de l'article téléchargé (donnée par la revue, mais très souvent différente de l'année où l'article a été mis en ligne sur le site d'Érudit)	1970	Dans les logs
Année de mise en ligne de l'article sur le site web d'Érudit (pas toujours possible à calculer)	1970	Calculé lors de la création de la base de données
Volume de la revue dans lequel se trouve l'article téléchargé	v24	Dans les logs
Numéro de revue dans lequel se trouve l'article téléchargé	n3	Dans les logs
Identifiant unique de l'article créé par Érudit et que l'on retrouve dans l'URL du téléchargement	302986	Dans les logs
L'article téléchargé provient-il d'une revue en libre accès complet, c.-à-d. sans embargo? (vrai ou faux)	f	Calculé lors du traitement des données

Annexe 3 : Structure de la base de données relationnelle conçue pour le projet

Tables	Champs		
Erudit_Logs.dbo.Download	ID	Continent	Referer_Host
	Article_ID	Country	OS
	Time	Region	Device_Type
	Local_Time	City	Age
	Proxy_IP	Timezone	Embargo
	User_IP	Browser	
Erudit_Logs.dbo.Article	ID		
	Issue_ID		
	Article (identifiant d'Érudit)		
Erudit_Logs.dbo.Issue	ID		
	Volume_ID		
	Issue		
	Publication_Year		
	Online_Year		
Erudit_Logs.dbo.Volume	ID		
	Journal_ID		
	Volume		
Erudit_Logs.dbo.Journal	ID		
	Journal (identifiant du référentiel de revues)		
	General_Discipline_Fr		
	Discipline_Fr		
	Speciality_Fr		
	Full_OpenAccess		
Erudit_Logs.dbo.Erudit_Articles	Article (identifiant d'Érudit)		
	Journal (identifiant du référentiel de revues)		
	Journal_Title		
	Issue		
	Volume		
	Publication_Year		
	Collection_Year (chaîne de caractères)		

Annexe 4 : Hiérarchie des techniques de détection des robots selon Doran et Gokhale (2010)

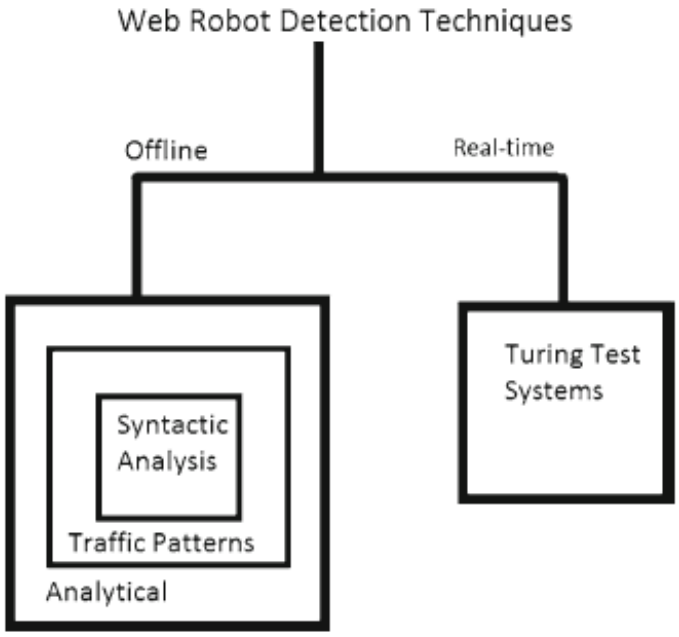

Fig. X – Hiérarchie des techniques de détection des robots (Doran et Gokhale, 2010, p. 189)